12th INTERNATIONAL SCIENTIFIC CONFERENCE ON DEFENSIVE TECHNOLOGIES
OTEH 2026
Serbia, Tara, 21 - 24 October 2026

METHODS FOR IMPROVING MACHINE LEARNING MODELS FOR PREDICTING MULTI-HOLE PRESSURE PROBE CHARACTERISTICS

IVANA PETKOVIĆ
Military Technical Institute, Belgrade, ivana.petkovic.vti@gmail.com

SNEŽANA ZUROVAC
Military Technical Institute, Belgrade, zurovac@medianis.net

DIJANA DAMLJANOVIĆ
Military Technical Institute, Belgrade, didamlj@gmail.com

ILIJA NENADIĆ
Military Technical Institute, Belgrade, ilijanenadic98@gmail.com

Abstract: *The standard engineering process for developing a multi-hole pressure probe included the use of 3D software during the design phase. At the same time, CFD (Computational Fluid Dynamics) was used to calculate the main characteristics of the probe. After the probe geometry was selected, the CFD results were verified by wind tunnel testing. This process ensured that the predicted probe characteristics were accurate under real flow conditions.*

However, numerical simulations and experimental testing required a significant amount of time and financial resources. For this reason, a machine learning model for predicting the characteristics of a multi-hole pressure probe was developed at the Military Technical Institute in Belgrade [1]. The selected model was a Random Forest Regressor because it achieved the lowest MAE (Mean Absolute Error).

The model was trained to predict probe characteristics for subsonic, transonic, and supersonic flow regimes and for two probe geometries: a sharp probe and a blunt probe. In the previous study, the Random Forest model achieved an MAE of 0,001681.

The aim of this research was to improve the performance of the existing model by using machine learning optimization methods. In addition, the effect of the dataset on model performance was analyzed. Grid Search with cross-validation was applied to find the best hyperparameter combination and improve the prediction accuracy of the model for different flow regimes.

Keywords: *multi-hole pressure probe, machine learning, Random Forest Regressor, hyperparameter optimization, flow regimes.*

1. INTRODUCTION

In previous research at the Military Technical Institute in Belgrade, a machine learning model based on the Random Forest Regressor was developed to predict the characteristics of a multi-hole pressure probe. The model was trained using experimental data from wind tunnel tests. The input data included flow parameters and probe geometry characteristics, while the output was ΔP/P0 (the normalized pressure difference) [1].

The previous model achieved a MAE of 0,001681. The Random Forest algorithm was selected as the best model after it was compared with other machine learning models [1]. Although the model showed high prediction accuracy, its performance could still be improved.

Since the performance of a machine learning model depended on the selected hyperparameters, this paper investigated different methods for improving the existing Random Forest model. The main focus was on hyperparameter optimization using Grid Search together with cross-validation to find the best combination of hyperparameters that gave the lowest prediction error.

In addition, the effect of dataset segmentation and dataset characteristics on model performance was analyzed. The analysis included different flow regimes (subsonic, transonic, and supersonic) and two different probe geometries. The results showed that model accuracy depended not only on the number of samples but also on the data distribution and homogeneity within each flow regime.

The aim of this paper was to improve the existing Random Forest model through hyperparameter optimization and analysis of the dataset structure.

Corresponding author: I. Petković;
E-mail: ivana.petkovic.vti@gmail.com:
DOI: 10.5937/OTEH26063P

2. DATASET DESCRIPTION

The dataset used in this study contained a total of 2095 samples collected for two probe geometries under different flow regimes. The samples were collected during wind tunnel tests in the T-35 and T-38 wind tunnels at the Military Technical Institute. The sharp probe is shown in Figure 1, while the blunt probe is shown in Figure 2. Each sample represented one measurement and included the following input parameters: static pressure (Pst), Mach number (M), flow temperature (T0), and effective angle of attack (θ_eff). The output variable was the normalized pressure difference (ΔP/P0).

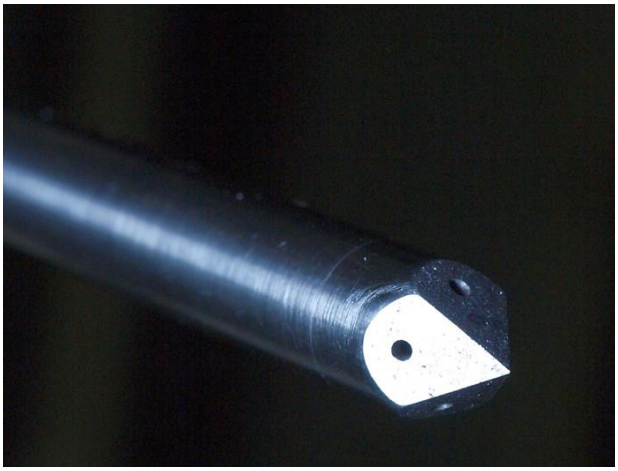


Figure 1. Four-hole pyramidal pressure probe, declared as *a sharp probe*; tested in the VTI T-38 wind tunnel in the Mach number range from 0.2 to 4

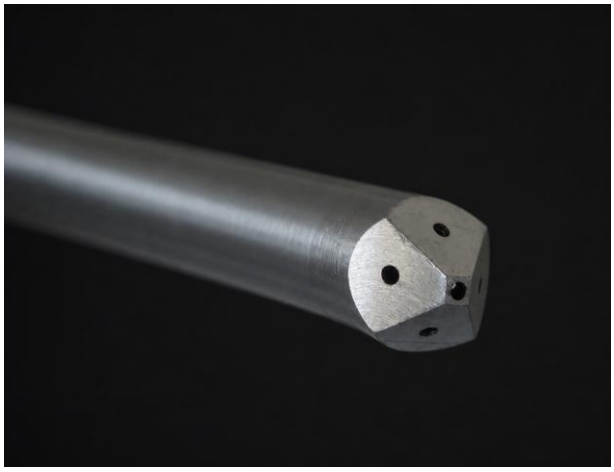


Figure 2. Five-hole pyramidal pressure probe designed for small UAV (unmanned aerial vehicle), declared as *a blunt probe*; single tested in the VTI T-38 wind tunnel

Data for the sharp probe were available for all three flow regimes (subsonic, transonic, and supersonic), while data for the blunt probe were available only for the subsonic flow regime. To better analyze the effect of the flow regime and probe geometry on model performance, the complete dataset was divided into four subsets. The number of samples in each subset is shown in Table 1.

The largest subset contained 1025 samples for the sharp probe in the supersonic flow regime, while the smallest subset contained 84 samples for the sharp probe in the transonic flow regime.

Table 1. Dataset structure

Dataset	Flow Regime	Geometry	Samples
sharp_subsonic	subsonic	sharp	672
sharp_transonic	transonic	sharp	84
sharp_supersonic	supersonic	sharp	1025
blunt_subsonic	subsonic	blunt	314

3. METHODOLOGY

3.1 Random Forest Regressor

In the previous study, the Random Forest Regressor was selected as the best model for predicting the characteristics of the pressure probe because it achieved the lowest MAE [1]. Based on these results, the same model was used in this study as the starting point for further optimization.

Random Forest is an ensemble machine learning method that consists of many decision trees. During training, each tree is built using a different subset of the training data. The final prediction is obtained by averaging the predictions of all trees. This approach reduces overfitting and improves the model's ability to make accurate predictions on new data [2].

To explain how the Random Forest algorithm works, Figure 3 shows one decision tree generated using the complete training dataset from this study. The maximum depth of the tree is limited to 3.

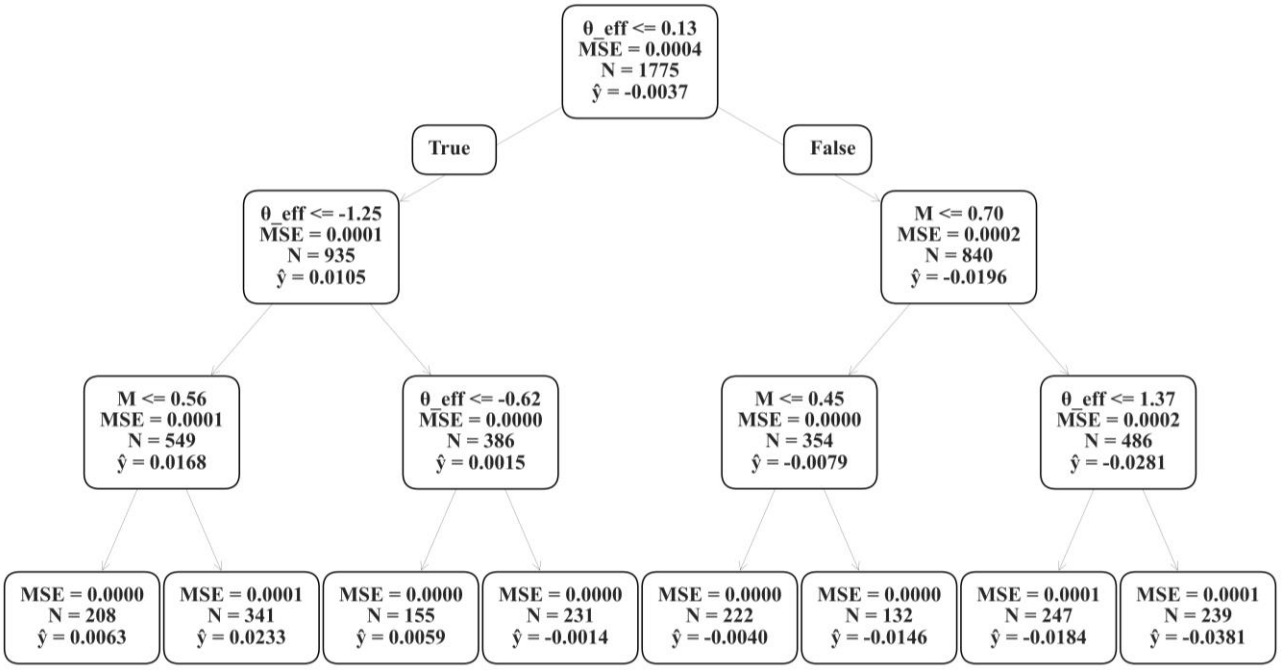


Figure 3. Example of a decision tree generated for the complete dataset

The root node represents the first data split. In this example, the algorithm selected the effective angle of attack, θ_eff, as the first splitting criterion. The threshold θ_eff ≤ 0.13° was automatically selected during training because it produced the largest reduction in the Mean Squared Error (MSE) among all possible splits of all input features. To find the best split, the algorithm checks different threshold values from the training data for each input feature. For every possible split, it calculates the MSE of the two new groups. The feature and threshold that give the lowest total MSE are selected.

All training samples are divided into two groups depending on whether θ_eff ≤ 0.13° or θ_eff > 0.13°. The root node contains all training samples used to train the model.

After the first split, the algorithm repeats the same process for each new group of data. At every node, it checks all input features and different threshold values again. The split that gives the largest reduction in MSE is selected. In this example, the next splitting criteria are again θ_eff and the Mach number (M), but with different threshold values. For example, on the right branch the next split is M ≤ 0.70, while on the left branch θ_eff is selected again with the threshold −1.25°. This shows that the algorithm chooses the best feature and threshold separately at each node.

Each node contains the Mean Squared Error (MSE), the number of training samples (N), and the predicted mean value of the target variable (ŷ). The parameter N represents the number of training samples contained in the node, while ŷ represents the average value of the target variable ΔP/P0 for those samples. The MSE shows how similar the target values are within a node. Lower MSE values indicate that the target values are more similar, meaning that the node is more homogeneous.

The splitting process ends at the leaf nodes. Each leaf contains training samples with similar input values. The prediction of a single decision tree is the value **ŷ**, which represents the average value of the target variable ΔP/P0 for all training samples contained in that leaf. The final prediction of the Random Forest Regressor is obtained by averaging the predictions of all decision trees in the forest.

3.2 Hyperparameter optimization

To improve the performance of the existing Random Forest model, hyperparameter optimization was performed using the Grid Search method. Grid Search tested all predefined combinations of hyperparameters to find the configuration that gave the lowest MAE [3].

In this study, the effect of three main hyperparameters of the Random Forest algorithm was analyzed: the number of trees in the forest (n_estimators), the maximum tree depth (max_depth), and the minimum number of samples required to split a node (min_samples_split). The tested hyperparameter values are shown in Table 2.

The selected hyperparameter values and different cross-validation configurations were combined in ten experiments to find the best model configuration. The cross-validation procedure used in this study is described in the next section.

Table 2. Tested hyperparameters

Parameter	Values
n_estimators	200, 300, 500, 700, 800, 1000
max_depth	10, 20, 30, 40, None
min_samples_split	2, 5, 7, 9, 10

3.3 Cross-validation

The performance of the model was evaluated using k-fold cross-validation. In this method, the dataset is divided into k subsets of approximately equal size. During the validation process, the model is trained using k−1 subsets, while the remaining subset is used for testing. This process is repeated k times, so that each subset is used once as the validation set [4].

In this study, cross-validation was used as part of the Grid Search procedure to select the best combination of hyperparameters for the Random Forest Regressor. In addition to hyperparameter optimization, the effect of the number of validation folds on model performance was also analyzed. For this purpose, ten experiments were performed using different values of the cv parameter (5, 10, 15, 40, 50, 60 and 75).

For each combination of hyperparameters, the average MAE was calculated over all validation iterations. The configuration with the lowest MAE was selected as the best model configuration.

3.4 Dataset segmentation

To study the effect of the dataset on the performance of the Random Forest regressor, the complete dataset was divided into four separate subsets. The division was based on the probe geometry and the flow regime. The four subsets were sharp subsonic, sharp transonic, sharp supersonic, and blunt subsonic.

The same Random Forest model was used for all four subsets. The model used the optimal hyperparameters found during the optimization process. The model was evaluated using k-fold cross-validation. This made it possible to compare the results under the same conditions.

The performance of each model was measured using the MAE. The Mach number distribution was analyzed for the sharp supersonic and blunt subsonic datasets because they had the highest and the lowest MAE. The distribution of the target variable ΔP/P0 was analyzed for all four dataset subsets using boxplot diagrams. The range and the standard deviation (Std) of the target variable were also calculated to measure the data variability.

The goal of this analysis was to find out whether the prediction accuracy depended only on the number of samples or also on the data distribution and variability.

4. RESULTS

4.1 Hyperparameter optimization results

The results of the hyperparameter optimization are shown in Table 3. Compared with the previously developed model [1], the optimized Random Forest Regressor achieved a lower MAE. The best result was obtained with the following hyperparameter values: n_estimators = 500, max_depth = 10, min_samples_split = 2, and cv = 50.

The obtained MAE was 0,001601, which represented an improvement of about 4.8% compared to the previous study, where the MAE was 0,001681. The results showed that hyperparameter optimization improved the prediction accuracy of the model, although the improvement was relatively small. This indicated that the original model already had high accuracy, while the optimization further improved its performance. All ten experiments were performed on the complete dataset.

Table 3. Comparison of MAE before and after hyperparameter optimization

Model	MAE
Random Forest (previous work)	0,001681
Random Forest (Grid Search optimized)	0,001601

4.2 Dataset segmentation results

The prediction results for the four dataset subsets are shown in Table 4. Each subset was analyzed separately using the Random Forest model with the same optimal hyperparameters (n_estimators = 500, max_depth = 10, and min_samples_split = 2). The model was evaluated using 50-fold cross-validation.

Table 4. Model performance for different dataset subsets

Dataset	Samples	MAE
sharp_subsonic	672	0,000804
sharp_transonic	84	0,001565
sharp_supersonic	1025	0,002799
blunt_subsonic	314	0,000405

The lowest MAE was obtained for the blunt subsonic dataset and was 0,000405. The highest MAE was obtained for the sharp supersonic dataset and was 0,002799. Although the sharp supersonic dataset contained the largest number of samples, it produced the highest prediction error. On the other hand, the sharp transonic dataset contained only 84 samples, but its MAE (0,001565) was much lower than the MAE of the sharp supersonic dataset. These results showed that a larger number of samples did not always lead to better prediction accuracy.

The distribution of the Mach number for the datasets with the highest and the lowest MAE is shown in Figure 4. It can be seen that the sharp supersonic dataset covered a much wider Mach number range than the blunt subsonic dataset. This indicated a larger variation of flow conditions in the sharp supersonic dataset.

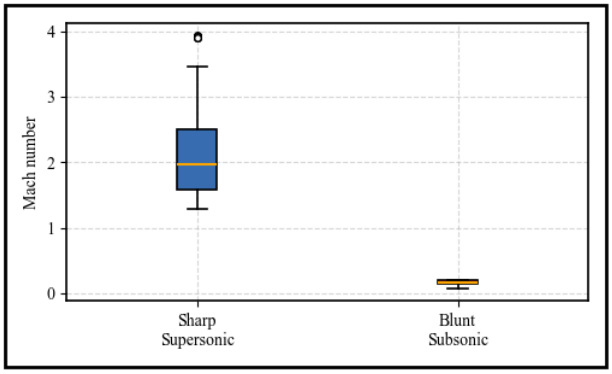


Figure 4. Distribution of Mach Number

Figure 5 shows the distribution of the target variable ΔP/P0 for all four dataset subsets. The differences between the distributions showed that the datasets had different levels of data variability.

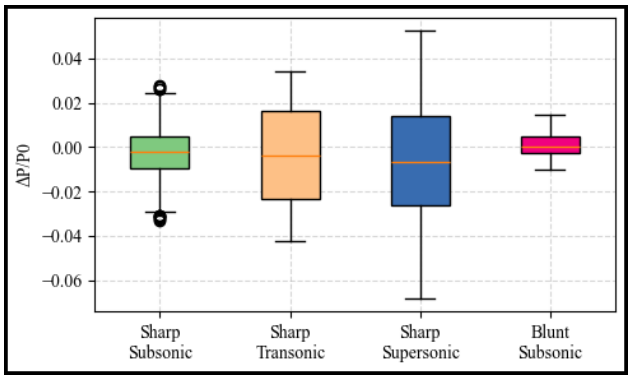


Figure 5. Distribution of ΔP/P0

For a more detailed analysis, the range, Std and MAE were calculated for each dataset subset. The results are shown in Table 5.

Table 5. Statistical characteristics of the dataset subsets

Dataset	Samples	Range	Std	MAE
sharp_subsonic	672	0,0614	0,0119	0,000804
sharp_transonic	84	0,0764	0,0225	0,001565
sharp_supersonic	1025	0,1203	0,0259	0,002799
blunt_subsonic	314	0,0249	0,0052	0,000405

The sharp supersonic dataset had the largest range (0,1203), the highest standard deviation (0,0259), and the highest MAE (0,002799). In contrast, the blunt subsonic dataset had the smallest range (0,0249), the lowest standard deviation (0,0052), and the lowest MAE (0,000405).

The results showed that the prediction accuracy of the Random Forest model did not depend only on the number of samples. Higher data variability and a wider data distribution made the learning process more difficult and increased the prediction error. On the other hand, more homogeneous datasets allowed the model to make more accurate predictions and achieve lower MAE values.

5. CONCLUSION

This paper analyzed different methods for improving the Random Forest Regressor developed for predicting the characteristics of a multi-hole pressure probe. By using Grid Search optimization and different cross-validation configurations, the MAE was reduced by about 4.8%. The best results were achieved with the following hyperparameter values: n_estimators = 500, max_depth = 10, min_samples_split = 2, and cv = 50.

In addition to hyperparameter optimization, the effect of dataset characteristics on model performance was also analyzed. The results showed that the sharp supersonic dataset, although it contained the largest number of samples, had the highest MAE and the highest variability of the target variable. In contrast, the blunt subsonic dataset had the lowest variability and the lowest MAE.

The results showed that, besides the number of samples, the data distribution and variability also had a significant effect on the accuracy of the Random Forest Regressor. They also showed that the structure of the dataset was an important factor that affected model performance. Therefore, future research will focus on a more detailed analysis of dataset characteristics and organization, as well as on developing methods for creating more representative training datasets.

Acknowledgement

This work was supported by the Ministry of science, technological development and innovations (Serbia), Contract No. 451-03-34/2026-03/200325.

References

[1] Zurovac S., Damljanović D., Nenadić I., Vuković Đ.: *Development of a Predictive Model for Aerodynamic Probe Design Based on Wind Tunnel Test Data*, 60th 3AF International Conference on Applied Aerodynamics, Paris - France, 2026.

[2] Breiman L.: *Random Forests*, Machine Learning, Vol. 45, No. 1, pp. 5–32, 2001.

[3] Pedregosa F., Varoquaux G., Gramfort A., et al.: *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830, 2011.

[4] Kohavi R.: *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1143, 1995.