RSM vs Machine Learning for moisture prediction in the convective drying of avocado pulp: Insights from a 30-run CCD Datasets

Awang Bono[1], Zykamilia Kamin[2#], Dona Stacy Petrus[2], Muhamad Afif Naqiudien Aladin[2], Mohd Hardyianto Vai Bahrun[3]

1 Faculty of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, Padang Besar 02100, Perlis, MALAYSIA.
2 Oil and Gas Engineering Programme, Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, MALAYSIA.
3 Chemical Engineering Programme, Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, MALAYSIA.
Corresponding author E-Mail: zykamilia@ums.edu.my; Tel: +60168315194.

ABSTRACT The amount of moisture content in drying process is crucial. Determining accurate process control, product quality and energy efficiency in food processing is necessary to ensure the quality of the drying food. In this study, Response Surface Methodology (RSM), a feedforward Neural Network (NN), and a Random Forest (RF) model were developed for moisture content of avocado pulp dried under controlled convective conditions. A Central Composite Design with number of experiments = 30 was used to explore four factors which are hot-air temperature (denoted as A, ranging from 40 to 70 °C), drying time (B, 13–20 h), wind speed (C, 0.12–0.26 m/s), and raw material thickness (D, 0.5–1.25 cm). A second-order polynomial model which was developed using RSM, was fitted to the experimental data and compared with other models, NN and RF using cross-validation to improve robustness. The results showed that RSM model performed significantly better than the other two models. The RSM model achieved coefficient of determination of 0.80 followed by NN, 0.43 and RF, 0.34. Feature-importance was used on ML models to determine the ranking of the factors based on how significant the factors influence the drying process. It was identified that wind speed and drying time as the main factors to represent the final moisture content, directly affect mass-transfer control in this application. The results highlight that small datasets can outperform the machine learning model especially for characterizing food drying processes.

KEYWORDS: Drying kinetics; Response Surface Methodology; Neural Network; Random Forest; Moisture prediction
Received 12 January 2026 **Revised** 27 March 2026 **Accepted** 1 April 2026 **In press** 1 April 2026 **Online** 2 April 2026

Original Article

TRANSACTIONS ON SCIENCE AND TECHNOLOGY

INTRODUCTION

Drying is one of the most critical complex processes in food processing, particularly for high-moisture fruits such as avocado. It reduces microbial activity and improving transportability of exposable products thus extending shelf life significantly (Başlar *et al.*, 2015). Some of the factors influencing the drying process are air temperature, air velocity, exposure duration, and material thickness (Aksoy *et al.*, 2019). The drying process is crucial because it determines the rate of moisture removal, quality retention and energy consumption of dried food. That is why it is important to understand how prediction of moisture content under different drying conditions affects process control, energy optimization, and high product quality (Delfiya *et al.*, 2022).

Traditionally, empirical and semi-empirical statistical methods such as Response Surface Methodology (RSM) have been widely applied in the optimization of food drying processes. RSM typically uses experiments results to build mathematical models that describe certain processes. Central Composite Designs (CCD) is one of the methods in experimental space design. Its advantage is it can reduce the number of experimental runs required while maintaining the quality of the outputs (Myers *et al.*, 2016). However, even RSM always has advantage to be interpretable and parsimonious models, it often assumes a quadratic relationship between factors and responses, which may not fully capture the nonlinear dynamics that exist in mass and heat transfer during drying.

In recent years, machine learning (ML) algorithms have become powerful alternatives for modelling complex nonlinear systems in food engineering. Some of the well-known algorithms, Artificial Neural Networks (ANNs), Random Forests (RFs), Support Vector Machines (SVMs), and Gradient Boosting Trees (GBTs) have demonstrated the ability to learn complex patterns not only in drying kinetics but also in other food processing (Fan *et al.*, 2025; Golpour *et al.*, 2020). Not only that, but these models are also built with approximation of nonlinear functions without being told. This means that they can prespecified equations and can integrate multiple affected factors simultaneously. Consequently, researchers from various previous studies used ML to optimize drying temperature, time, and air velocity for various agricultural products to provide accurate predictions of moisture ratio, drying rate, and energy efficiency (Khan *et al.*, 2022; Rather *et al.*, 2024).

Nevertheless, the effectiveness of ML depends strongly on the size and quality of available data. Typically, designing an experiment in RSM requires small amount of dataset which is around 20-40. This small number of datasets limits the ability of ML model to learn the intricate pattern between the independent and dependent variables. Under such circumstances, complex models like neural networks or ensemble methods can underfit because it does not have enough data to learn to capture meaningful patterns, therefore perform poorly (Zantvoort *et al.*, 2024). To avoid this issue, cross-validation (CV) techniques are essential. In particular, k-fold cross-validation randomly separate and arrange the small dataset into k equally smaller subsets of the datasets. This process minimizes the overfitting issue by providing a more reliable estimate of model robustness and helps select optimal model complexity (Refaeilzadeh *et al.*, 2018).

Most earlier studies already used RSM to study drying behaviour. However, they usually focused on only using RSM despite the popularity of ML algorithms nowadays. In this work, RSM is compared directly with machine learning models. Additionally, instead of only checking how well the model fits, cross-validation is used to see how stable the models are. This study also adds feature importance for ML models which linked to the drying process; thus, it is not just numbers but also has physical meaning. Overall, this study shows that a simple model like RSM can combine with ML models when the data is limited which is something that rarely reported in previous RSM studies. In this study, a dataset from a previous published article was reused to conduct a secondary, independent modelling comparison using Response Surface Methodology (RSM), feedforward Neural Network (NN), and Random Forest (RF) model for the prediction of final moisture content in avocado pulp dried under controlled convective conditions.

MATERIAL AND METHODS

Datasets Source and Experimental Design

The experimental data used for modelling were obtained from a study by Kowarit *et al.* (2024), which investigated convective drying of avocado pulp under varied operational ranges (Kowarit *et al.*, 2024). The dataset was reused and reanalysed in this work to perform comparative modelling using Response Surface Methodology (RSM), Neural Network (NN), and Random Forest (RF) algorithms. This is to ensure the consistency of the physical data while enabling evaluation of different predictive models under identical experimental conditions.

A Central Composite Design (CCD) was employed to examine four process parameters. Thirty experiments were conducted, including factorial, axial, and center points with alpha value, $\alpha = 2.0$. The experiments were performed using avocado pulp of uniform size and initial moisture content,

and moisture content was measured gravimetrically. Table 1 summarizes the experimental design conducted in this work.

Table 1. Summary of Central Composite Design (CCD) factor ranges.

Factor	Description	Range
A	Hot Air Temperature (°C)	40 – 70
B	Drying Time (h)	13 – 20
C	Wind Speed (m/s)	0.12 – 0.26
D	Pulp Thickness (cm)	0.5 – 1.0

Modeling Approaches

Three predictive modelling approaches were implemented: RSM, feedforward Neural Network (NN), and Random Forest (RF). The goal of this approach is to develop empirical and data-driven models that can estimate the final moisture content (Y) as a function of the four process parameters (A, B, C, D). All variables were treated in actual form during regression.

Response Surface Methodology (RSM)

Second-order polynomial model predicting moisture content as a function of process variables is shown in (Equation 1) below:

$$\begin{aligned}\text{Moisture Content} &= 18.9 + 0.15375A + 0.727083B + 0.26375C - 0.0695833D - 0.184375AB \\ &+ 0.403125AC - 0.398125AD + 0.555625BC - 0.298125BD - 0.595625CD \\ &- 0.283437A^2 + 0.329063B^2 + 0.641563C^2 + 0.742812D^2\end{aligned} \quad (1)$$

where the squared terms represent factors interaction effects of themselves. This equation is developed using identical RSM-CCD analysis from the original article by Kowarit *et al.* (2024). The equation includes linear, interaction, and quadratic terms of the four independent variables. Fitting of the model was performed using least-squares regression, and the significance of model based on the coefficients value was evaluated through ANOVA analysis. The adequacy of the RSM model was verified using coefficient of determination (R^2).

Machine Learning Models

The machine learning models were implemented in Python (version 3.10) using the scikit-learn library, which provides a consistent and complete framework for regression and model evaluation (Pedregosa *et al.*, 2011). MinMaxScaler was used to normalize all input data to become more manageable for the models training. A simple feedforward Neural Network model was implemented using the MLPRegressor module in scikit-learn library. Hyperparameters were optimized using grid search within the cross-validation framework. The optimized neural network architecture has one hidden layer with eight neurons, rectified linear unit (ReLU) as activation function, and the Adam with adaptive learning as optimizer. For model training, 500 iterations were set with an adaptive learning rate. A fixed random-state seed of 42 was used for all models to ensure consistent results across runs.

The Random Forest model was developed using the RandomForestRegressor class with 100 estimators and a maximum tree depth of 5, balancing bias variance trade-off given the small dataset size. The model performs training predictions from multiple decision trees on bootstrap samples to

improve robustness against noise and overfitting. Feature importance was calculated from the mean decrease in impurity across all trees, making interpretation possible.

Cross-Validation and Model Evaluation

ML models were assessed using k-fold cross-validation (k = 5), meaning that five mutually exclusive subsets were partitioned for these models. In each iteration, four folds were used for training and one for testing, ensuring that each sample was used once for validation. The overall model performance was quantified by averaging the coefficient of determination (R^2). This approach avoids bias towards the small dataset samples, providing a more realistic measure of predictive reliability (Kuhn & Johnson, 2019; Refaeilzadeh *et al.*, 2018). The RSM model that differentiates from other two models, being an analytical regression model, was evaluated using the same data splits for direct comparability. Residual analysis was used to identify potential outliers that contribute to deviation of the predictions.

RESULTS AND DISCUSSION

Model Performance

The predictive performance of the three modeling approaches is shown below. Table 2 summarizes the average determination coefficients (R^2) for each model, reflecting the accuracy and generalization ability of their predictions.

Table 2. Model performance comparison based on R^2 values.

Model	R^2 Value	Observation
RSM	0.80	Accurate overall trend; moderate scatter
NN	0.43	Clustered predictions; narrow range (18–24%)
RF	0.34	Clustered predictions; narrow range (18–24%)

Based on the table, the RSM model achieved the highest predictive accuracy (R^2 = 0.80), indicating that the second-order polynomial has highest accuracy among the experimental data. In contrast, the NN and RF models exhibited significantly lower performance (R^2 = 0.43 and 0.34, respectively). These results are consistent with previous findings that data-driven algorithms tend to underfit small designed experimental datasets which were not enough for the ML models to learn (Zantvoort *et al.*, 2024). Although both NN and RF showed lower performance of predicted versus actual data points near the unity line, their predictions were narrow the output range (18–24% moisture), suggesting poor extrapolation beyond the mean. The cross-validation approach confirmed that this limitation was not due to random sampling variation but to structural overfitting, as the models failed to predict across other folds. Similar observations have been reported in another study where limited drying data limit the learning capability of ML models (Przybył & Koszela, 2023). The RSM model's superior performance highlights the advantage of low-parameter models in small data regimes. Since RSM explicitly determines the interaction and curvature effects from the values in polynomial terms, it efficiently captures the important trends without requiring large datasets for training like ML models (Myers *et al.*, 2016).

Cross-Validation Insights

Cross-validation was used to check how well the models generalize. The RSM model stayed consistent across all folds, with a standard deviation below 0.05, showing that its predictions were stable and not too sensitive to how the data was split. In contrast, the NN and RF models varied a little bit more, with deviations reaching up to ±0.15. This kind of variation suggests that NN and RF

tend to fit the training data too closely but struggle when tested on new data, which is a common issue with small datasets (Kuhn & Johnson, 2019). Overall, even if NN and RF look good during training, their ability to generalize is still limited, highlighting the need for careful validation in small-sample studies.

Feature Importance Interpretation

The analysis from feature-importance for both NN and RF models are shown in Figure 1(a) and (b), respectively. It was revealed that wind speed (C) and drying time (B) were the most significant factors influencing final moisture content. These findings are consistent with other drying study, which attributes moisture removal primarily to external convective mass transfer and exposure duration (Başlar *et al.*, 2015; Delfiya *et al.*, 2022). Air temperature (A) and material thickness (D) showed lower importance scores, indicating less significance effects within the studied range. The weaker temperature dependence suggests that, under the tested convective conditions, mass-transfer limitations were more significant than thermal driving forces. These results are in agreement with the other studies that investigated the characteristic behavior in the falling-rate period of drying (Aksoy *et al.*, 2019; Khan *et al.*, 2022). The trends confirm the consistency between data-driven (ML) and empirical models (polynomial model). This similarity supports the physical validity of the identified relationships and provides confidence in the interpretability among the models even though each of them has different mathematical formulations.

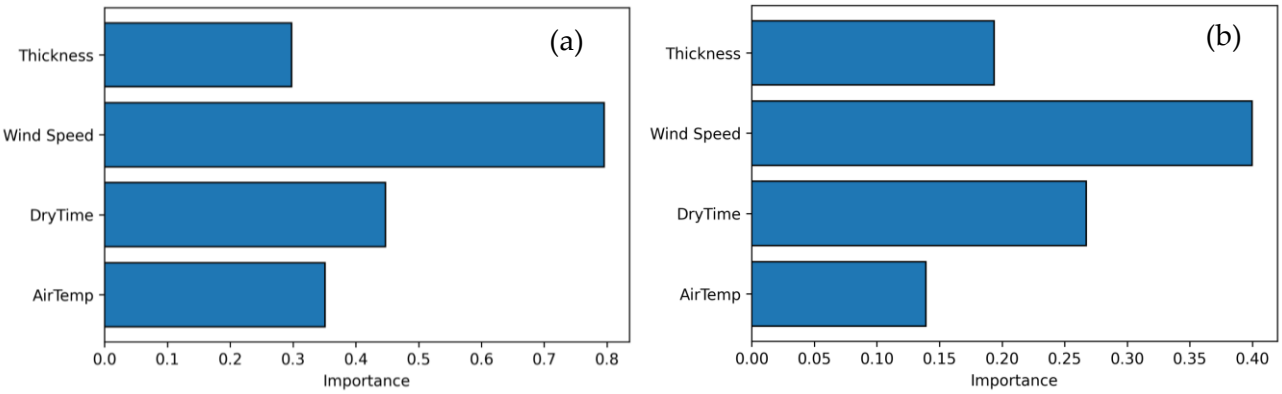

Figure 1. Feature importance determined by the (a) NN and (b) RF model.

Predicted vs. Actual Comparison

Figure 2(a), (b) and (c) illustrate the predicted versus actual moisture content for RSM, NN and RF model, respectively. The RSM model showed close predicted data points with experimental data, capturing the full observed range of roughly around 17–25% moisture. Comparison between predicted vs actual show**s** that the second-order RSM model accurately captures the full range of observed moisture values (17–25%), demonstrating strong linear agreement with experimental data. Moreover, each of the data points were distributed evenly along the unity line. By contrast, the NN and RF model generated more clustered predictions within the narrower 18–24% interval. The figures obviously show that almost no correlation exists between the predicted and actual data points. The data points are all over the place that makes the prediction difficult. Predictions appear tightly clustered along the unity line but are limited to a narrow range (18–24%), reflecting restricted generalization and possible overfitting around the region. The RF model exhibits a similar pattern to the NN model, with accurate predictions only within a limited subset of the observed data range. These results confirm that RSM retained better predictive flexibility across the variable domain, while ML models demonstrated restricted learning due to limited data coverage.

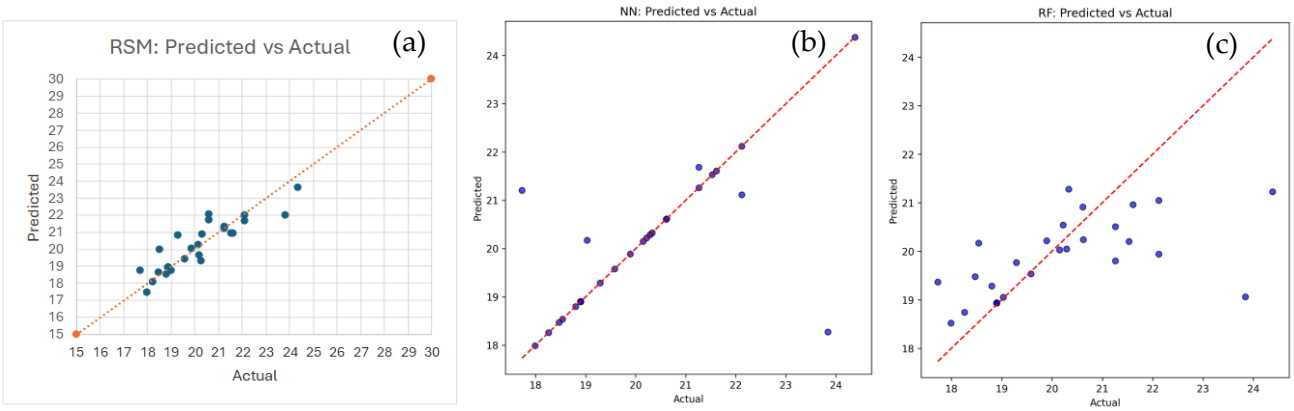

Figure 2. Predicted versus actual moisture content using the (a) RSM, (b) NN and (c) RF model.

Error Analysis

The error analysis for RSM, NN and RF are shown in Figure 3(a), (b) and (c), respectively. Each model was analyzed to identify runs that contribute to the overall false prediction. These high-residual runs may be due to human error from experiments, measurement uncertainty, or unmodeled interactions among process factors. For the RSM model, the residuals followed a more uniform distribution, supporting the adequacy of the quadratic model structure. Nonetheless, it is good to investigate the high-error points because it can provide valuable insights for refining future experimental designs or introducing hybrid correction terms. Such analysis is consistent with best practices among other researchers in model validation and iterative refinement in process modeling (Kuhn & Johnson, 2019). For NN model, a small number of experimental runs dominate total error, suggesting potential measurement anomalies or unmodeled variations. Conversely, the error distribution for RF model is heavily skewed, with several high-residual samples contributing disproportionately to overall prediction variance. These findings are consistent with the overall accuracy performance as reported in the earlier results of this study.

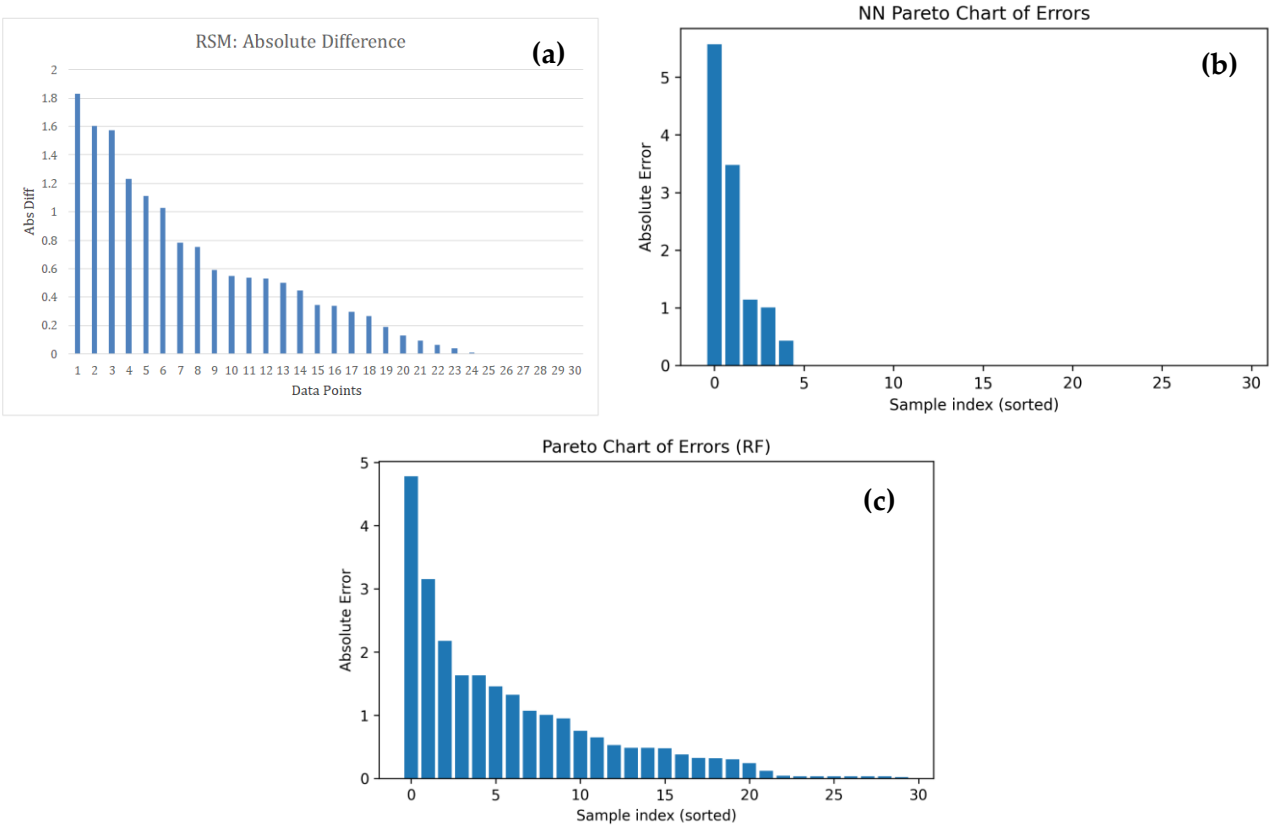

Figure 3. Pareto chart of absolute residuals for the (a) RSM, (b) NN, and (c) RF model.

CONCLUSION

Comparison predictive performance study has been shown by using three models, namely Response Surface Methodology (RSM), Neural Network (NN), and Random Forest (RF) to estimate the final moisture content of avocado pulp dried under controlled convective conditions. Among these models, RSM achieved highest accuracy, significantly outperforming both NN and RF. The better performance of RSM can be due to its mathematical model formulation while lower ML models performance directly learns from small dataset. Although ML models are inherently powerful for modeling nonlinear behavior, their predictive power depends on the volume and diversity of training data. ML model needs to learn data that contains large enough variations in the datasets to be able to capture the nonlinear complex patterns. This outcome proved the importance of using k-fold cross-validation in small number of dataset ML studies. Even though NN and RF models are weak in predictive capability, they still able to identify significance of parameters as they ranked drying time (B) and air velocity (C) as the most influential factors that control final moisture content, while air temperature (A) and material thickness (D) did not affect the final moisture content too much. Future work should focus on increasing the number of datasets. It can be either through additional experimental runs or integration of synthetic data generation techniques. Moreover, combining the interpretability of RSM with the nonlinear learning capabilities of ML such as RSM-guided neural networks or ensemble stacking may provide higher accuracy. Therefore, ML algorithms are expected to be high demand in predicting and optimizing food drying in the future because of their capability to easily learn complex patterns with the availability of datasets.

REFERENCES

[1] Aksoy, A., Karasu, S., Akcicek, A. & Kayacan, S. 2019. Effects of different drying methods on drying kinetics, microstructure, color, and the rehydration ratio of minced meat. *Foods*, *8*(6), 216.

[2] Başlar, M., Kılıçlı, M. & Yalınkılıç, B. 2015. Dehydration kinetics of salmon and trout fillets using ultrasonic vacuum drying as a novel technique. *Ultrasonics Sonochemistry*, *27*, 495–502.

[3] Delfiya, D. S. A., Prashob, K., Murali, S., Alfiya, P. V, Samuel, M. P. & Pandiselvam, R. 2022. Drying kinetics of food materials in infrared radiation drying: A review. *Journal of Food Process Engineering*, *45*(6), e13810.

[4] Fan, L., Pei, Y., Zhang, L., Kong, J. & Xu, W. 2025. Applications of Machine Learning Models in Agricultural Product Drying: A Comprehensive Review of Advances, Challenges, and Prospects. *Food and Bioprocess Technology*, 18, 10047–10085.

[5] Golpour, I., Kaveh, M., Amiri Chayjan, R. & Guiné, R. P. F. 2020. Optimization of infrared-convective drying of white mulberry fruit using response surface methodology and development of a predictive model through artificial neural network. *International Journal of Fruit Science*, *20*(sup2), S1015–S1035.

[6] Khan, M. I. H., Sablani, S. S., Joardder, M. U. H. & Karim, M. A. 2022. Application of machine learning-based approach in food drying: Opportunities and challenges. *Drying Technology*, *40*(6), 1051–1067.

[7] Kowarit, S., Sathapornprasath, K. & Jansri, S. N. 2024. Application of hot air-derived RSM conditions and shading for solar drying of avocado pulp and its properties. *Solar Energy*, *278*, 112768.

[8] Kuhn, M. & Johnson, K. 2019. *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.

[9] Myers, R. H., Montgomery, D. C. & Anderson-Cook, C. M. 2016. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (4th Ed). Wiley.

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *The*

Journal of Machine Learning Research, 12, 2825–2830.

[11] Przybył, K. & Koszela, K. 2023. Applications MLP and other methods in artificial intelligence of fruit and vegetable in convective and spray drying. *Applied Sciences*, 13(5), 2965.

[12] Rather, I. H., Kumar, S. & Gandomi, A. H. 2024. Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets. *Artificial Intelligence Review*, 57(9), 226.

[13] Refaeilzadeh, P., Tang, L. & Liu, H. 2018. Cross Validation. *In:* Liu, L. & Özsu, M.T. (eds). *Encyclopedia of Database Systems*. Springer. pp. 532–538.

[14] Zantvoort, K., Nacke, B., Görlich, D., Hornstein, S., Jacobi, C. & Funk, B. 2024. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digital Medicine*, *7*(1), 361.