Bias-Corrected Imputation and Metaheuristic Boosting for Multi-Output Prediction

Tidarat Luangrungruang
Department of Computer
Faculty of Science and Technology
Sakon Nakhon Rajabhat University
Sakon Nahon, Thailand
tidarat@snru.ac.th

Gawalee Phatai*
Department of Computer
Faculty of Science and Technology
Sakon Nakhon Rajabhat University
Sakon Nahon, Thailand
gawalee@snru.ac.th

***Abstract*— Missing values present a critical challenge in multi-output prediction, as they can propagate bias across outputs through shared representations. Conventional imputation techniques often overlook inter-variable dependencies and fail to reduce bias adequately. To address this issue, this study proposes a bias-corrected predictive imputation framework combined with metaheuristic-optimized boosting models. In this study, Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), which inherently handle missing values during training, were further enhanced through predictive imputation. Their performance was most notably improved when combined with the Teaching–Learning-Based Optimization (TLBO) algorithm, which provided significant reductions in error metrics and strengthened generalization. Experiments on the dataset demonstrate that predictive imputation enhances accuracy by effectively addressing missing values, while bias correction and TLBO optimization substantially reduce error metrics and improve generalization. LightGBM with TLBO achieved the best performance, exceeding 90% accuracy for both outputs. These results highlight the effectiveness of bias-corrected predictive imputation and metaheuristic-optimized boosting in delivering robust and accurate multi-output prediction.**

Keywords*— Bias correction, multi-output prediction, boosting models, predictive imputation, metaheuristic-optimization

I. INTRODUCTION

The challenges associated with large and complex datasets become more pronounced when developing multi-output prediction models [1], which require complete and high-quality data to accurately learn the relationships between input variables and multiple outputs simultaneously. However, missing values remain a critical issue that can reduce model learning efficiency and increase the bias of prediction results [2]. This issue is particularly severe when missing data are Not Missing at Random (NMAR), where the missingness depends on the true values of the variables [3]. In such cases, learning from an incomplete subset may propagate bias across multiple outputs, and joint loss functions, such as Mean Squared Error (MSE), may be evaluated incompletely.

Although simple imputation methods have been proposed, they are limited in multi-output settings because they ignore inter-output dependencies and cannot adequately reduce bias. Prior studies have focused either on missing-value handling or improving multi-output prediction performance, but few integrate predictive imputation with metaheuristic hyperparameter optimization [4][5].

To address this gap, this study proposes an approach combining predictive imputation with multi-output prediction models using XGBoost and LightGBM, with hyperparameters optimized via the Teaching–Learning-Based Optimization (TLBO) algorithm. The models are evaluated using multiple metrics, including MSE, MAE, MRE, and accuracy, to comprehensively assess improvements in predictive performance and bias mitigation.

II. METHODOLOGY

A. Bias Correction for Missing Data Imputation

In statistical data analysis, missing data is a common problem. If the model used for imputing missing values is misspecified, it may lead to biased parameter estimation [6]. To address this issue, bias correction can be applied after the imputation process. The main idea is to use an imputation model that allows for easy random sampling of missing values. After imputation, the likelihood of the model is adjusted using weights calculated from the density ratio between the true conditional density of the missing data and the conditional density from the imputation model. These weights are then applied to adjust the likelihood function used for parameter estimation, compensating for the effects of model misspecification in the imputation process. The advantage of this method is that it reduces the bias of the estimates without requiring specification of the missingness mechanism, which is often difficult in practice. Moreover, it can be combined with different imputation models [7][8].

B. Predictive Imputation (PI)

PI, or the imputation of missing values through prediction, is a method of handling missing data by using a model to estimate the missing values from other variables in the dataset [9]. Unlike simple imputation techniques such as mean, median, or mode substitution, which often disregard the relationships among variables, predictive imputation exploits existing dependencies within the data. The advantage of this approach is that it can capture both linear and nonlinear relationships, producing imputed values that are more plausible and closer to the true values. It also helps to reduce bias and variance in the data compared to simpler imputation methods. In practice, several approaches to predictive imputation are commonly used. For this study, single imputation with a Decision Tree Regressor was employed. A Decision Tree is a supervised learning algorithm that recursively partitions data based on feature values, allowing it to model complex and nonlinear relationships without prior statistical assumptions. This method imputes one column at a time by constructing a decision tree model from the remaining observed variables, without assuming prior statistical distributions (non-parametric). Such an approach is particularly suitable for complex datasets with nonlinear relationships. The main strength of this method lies in its flexibility and its ability to capture complex patterns in the data more effectively than conventional imputation techniques [10][11].

* corresponding author

C. *Extreme Gradient Boosting (XGBoost)*

XGBoost is an advanced machine learning algorithm derived from Gradient Boosted Decision Trees, designed for high computational efficiency and scalability [12]. As an ensemble method, it iteratively constructs multiple weak learners, where each learner focuses on correcting the residual errors from the previous iteration. One of its key advantages is the ability to handle missing values inherently during the training process. Instead of requiring prior imputation, XGBoost automatically learns the optimal direction to assign missing data in the decision tree branches. This capability reduces the risk of introducing additional bias from external imputation and enables the model to exploit the underlying data structure more effectively. Consequently, XGBoost is particularly robust in dealing with incomplete datasets while maintaining high predictive accuracy in both classification and regression tasks [13][14].

D. *Light Gradient-Boosting Machine (LightGBM)*

LightGBM is an algorithm developed from decision trees by incorporating an ensemble technique known as boosting [15]. In this approach, multiple weak decision tree models are sequentially combined, where each subsequent tree corrects the errors of the preceding one until an optimal model is obtained. This type of algorithm is referred to as Gradient Boosting Decision Trees (GBDTs), which provides an efficient framework for both classification and regression tasks. A distinctive feature of LightGBM is its inherent ability to handle missing values during the training process. Instead of requiring prior imputation, LightGBM automatically determines the optimal split direction for missing values within the decision tree construction. This mechanism reduces the dependency on external preprocessing steps and mitigates the risk of bias introduced by simple imputation methods. Combined with its leaf-wise growth strategy—which enables more effective loss reduction and faster computation—LightGBM demonstrates robust performance when applied to large-scale, high-dimensional, and incomplete datasets [16][17].

E. *Teaching–Learning-Based Optimization (TLBO)*

TLBO is a population-based metaheuristic inspired by the classroom learning process [18]. Each candidate solution represents a learner, with the best solution regarded as the teacher. The algorithm iterates through two phases. The Teacher Phase, where learners are guided toward the teacher's performance, and the Learner Phase, where learners improve through peer-to-peer interaction. This mechanism balances exploration and exploitation, progressively refining solutions [19][20]. Before applying TLBO, missing values in the dataset were first imputed using Predictive Imputation (PI), and the optimization was performed on the resulting complete dataset. By systematically evaluating different parameter configurations, TLBO identifies the combinations that yield the highest predictive performance [21][22]. This integration enhances the ability of XGBoost and LightGBM to generalize from data, reduces error metrics, and strengthens model robustness in multi-output prediction tasks, particularly in the presence of missing values.

III. Experiments and Results

In this study, two boosting algorithm models were employed: Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM). These models were selected due to their proven efficiency and strong predictive performance, particularly when working with large-scale, high-dimensional datasets. Both algorithms are well-suited for scenarios involving imputation preprocessing. The models were evaluated in combination with predictive imputation techniques and compared against each other, with their hyperparameters optimized using the metaheuristic algorithm. To further illustrate the proposed method, we conducted an analysis using the UCI Bias Correction of Numerical Prediction Model Temperature Forecast dataset [23], comprising 7,750 instances, 23 input features, and 2 output variables. This dataset supports the bias correction of next-day maximum and minimum temperature forecasts from the Local Data Assimilation and Prediction System (LDAPS) model over Seoul, South Korea, using summer data from 2013 to 2017. The inputs include LDAPS forecasts, observed temperatures, and geographic variables, while the outputs are the next-day maximum and minimum temperatures. Model evaluation was conducted using a 10-fold cross-validation scheme repeated 10 times to ensure the robustness and statistical reliability of the findings across all five models. In each repetition, the dataset was randomly shuffled and split into training and testing subsets. For every iteration, the mean absolute error (MAE), mean relative error (MRE), mean squared error (MSE), and accuracy (ACC) were calculated for both training and testing sets. All reported performance values correspond to the error measured on the testing, or unseen, data.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

$$MRE = 100 \times \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|)^2 \tag{3}$$

$$ACC = \frac{1}{n}\sum_{i=1}^{n}1(|y_i - \hat{y}_i| \le \epsilon) \times 100 \tag{4}$$

In the above equations, y_i denotes the true value of the output variable, $\hat{y}_i$ represents the predicted value, and n is the number of samples in the training or testing dataset, where i ranges from 1 to n.

For regression evaluation, ACC is defined as the proportion of predictions within a specified tolerance of the true values. Formally, a prediction $\hat{y}$ is considered accurate if its absolute error does not exceed a specified tolerance ϵ. In this study, the tolerance ϵ was set to 0.05 for all outputs, and ACC was computed as the percentage of predictions meeting this criterion.

Table I presents the predictive performance of all model configurations on unseen test data for the two output variables (Y1 and Y2). The models were evaluated using four metrics: MAE, MRE, MSE, and ACC.

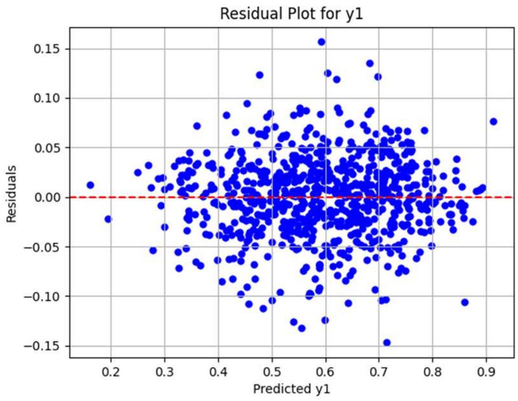
TABLE I. METRICS PERFORMANCE ON UNSEEN DATA FOR PREDICTING OF EACH

Model	Y1				Y2			
	MAE	MRE	MSE	ACC.	MAE	MRE	MSE	ACC
XGBoost	0.0320	0.0584	0.0017	79.22	0.0279	0.0489	0.0013	84.95
LightGBM	0.0321	0.0586	0.0017	79.18	0.0279	0.0488	0.0013	85.03
XGBoost with PI	0.0313	0.0572	0.0017	80.2	0.0275	0.0484	0.0013	85.46
LightGBM with PI	0.0317	0.0582	0.0017	79.68	0.0279	0.0490	0.0013	85.12
XGBoost with TLBO	0.0261	0.0478	0.0012	86.94	0.0235	0.0416	0.0010	90.30
LightGBM with TLBO	0.0257	0.0468	0.0012	87.76	0.0232	0.0410	0.0010	90.64

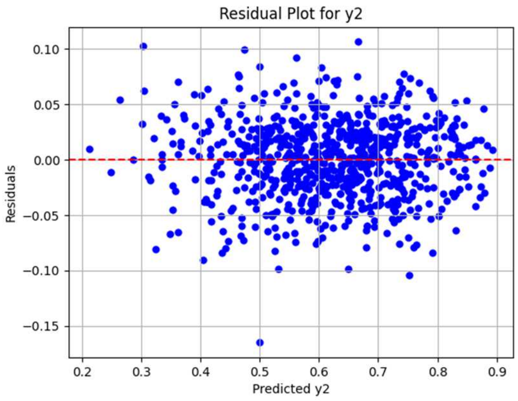
For the baseline models without missing value handling, XGBoost and LightGBM achieved similar performance, with MAE values around 0.032 for Y1 and 0.028 for Y2, and accuracy levels ranging from 79.18% to 85.03%. Incorporating PI led to modest improvements: XGBoost with PI reduced MAE and MRE while slightly increasing ACC, whereas LightGBM with PI showed minor improvements for Y1.

The largest performance gains were observed when the models were optimized using TLBO. XGBoost with TLBO reduced the MAE to 0.0261 for Y1 and 0.0235 for Y2, with corresponding accuracies of 86.94% and 90.30%. LightGBM with TLBO further improved performance, achieving the lowest MAE, MRE, and MSE values, as well as the highest ACC for both outputs (87.76 for Y1 and 90.64 for Y2). These results indicate that combining predictive imputation with metaheuristic optimization significantly enhances both the accuracy and stability of XGBoost and LightGBM for multi-output prediction tasks.

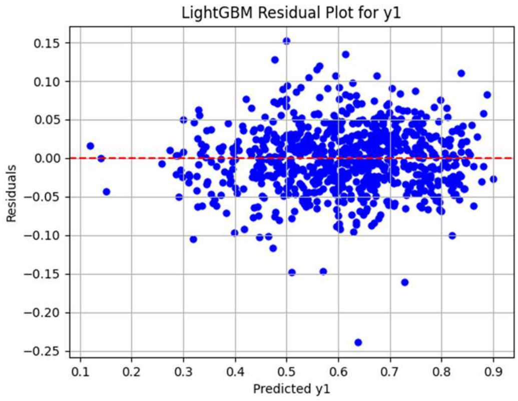
Residual plots for all models and both output variables are shown in Figure 1. Residual analysis was performed to evaluate the validity with particular attention to whether the errors were randomly distributed around zero and showed no identifiable systematic patterns.



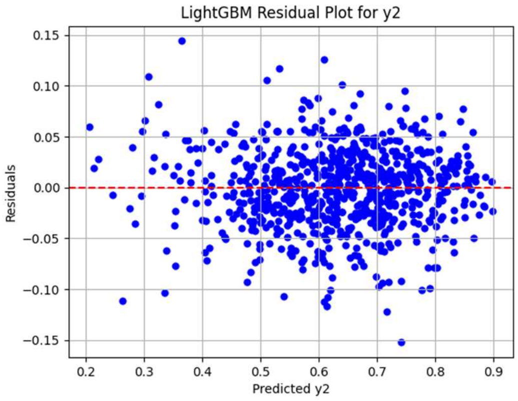
XGboost Y1



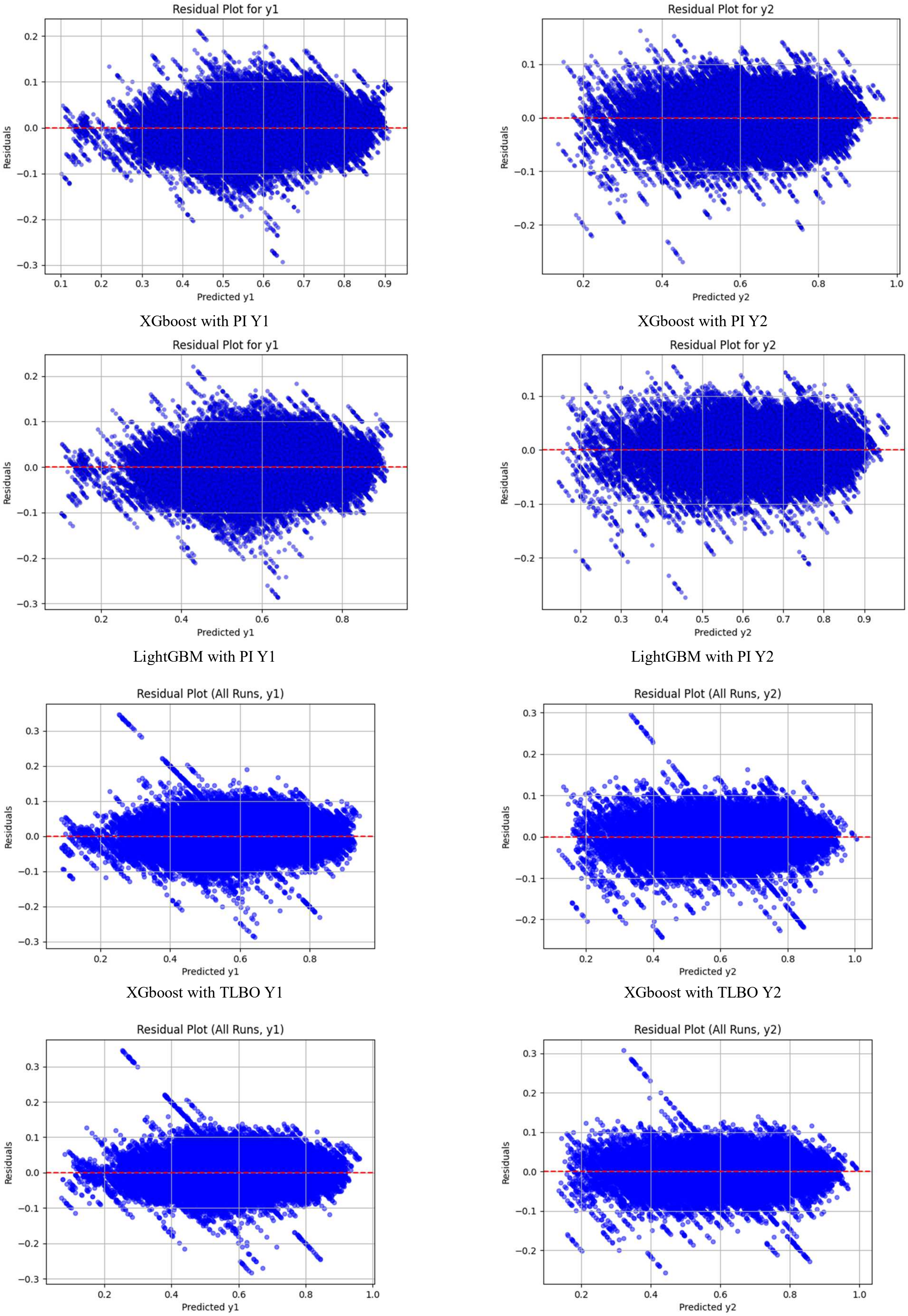
XGboost Y2



LightGBM Y1



LightGBM Y2



XGboost with PI Y1

XGboost with PI Y2

LightGBM with PI Y1

LightGBM with PI Y2

XGboost with TLBO Y1

XGboost with TLBO Y2

LightGBM with TLBO Y1

LightGBM with TLBO Y2

Fig. 1. The residual plots of all models.

The residual plots of all models for the two output variables are presented in Figure 1. Residual analysis was conducted to evaluate whether the errors were randomly dispersed around zero without discernible systematic patterns. XGBoost reduced some of these patterns but still exhibited clustered residuals in certain regions of the predicted values. In contrast, LightGBM displayed residuals more evenly distributed around the zero line, suggesting a stronger ability to capture nonlinear relationships in the data and mitigate systematic bias.

Further improvements were observed when predictive imputation was incorporated into both models. Specifically, XGBoost with predictive imputation and LightGBM with predictive imputation demonstrated more balanced residual distributions and reduced clustering effects. Moreover, the integration of the TLBO algorithm further enhanced model performance. In particular, XGBoost with TLBO and LightGBM with TLBO showed the most notable improvements, with residuals appearing more randomly dispersed and centered around zero, indicating better generalization and reduced systematic error. These results highlight the effectiveness of imputation and optimization techniques in improving model robustness and predictive accuracy.

IV. Conclusion

This study introduced a bias-corrected predictive imputation framework integrated with metaheuristic-optimized boosting models to address the persistent challenge of missing values in multi-output prediction. While XGBoost and LightGBM inherently provide mechanisms to handle missing values during training, experimental results demonstrate that their performance can be substantially improved through predictive imputation. By explicitly modeling inter-variable dependencies, predictive imputation reduces bias introduced by incomplete data and enhances the plausibility of imputed values. The integration of TLBO further reinforced the models, delivering significant reductions in error metrics and improving generalization performance across both outputs. Among all configurations, LightGBM with TLBO achieved the highest predictive accuracy, exceeding 90% for both maximum and minimum temperature forecasts, thereby validating the effectiveness of the proposed framework.

The findings highlight several important contributions. First, the study provides evidence that predictive imputation is not only complementary but also synergistic with boosting algorithms that already incorporate missing value handling mechanisms. Second, the integration of metaheuristic optimization with boosting models offers a systematic approach to parameter tuning, leading to consistent improvements in predictive performance. Third, the application to the dataset demonstrates the framework's potential for real-world forecasting tasks where missing values and multi-output relationships are unavoidable. Collectively, these contributions address a key research gap by showing that the joint consideration of bias correction, imputation, and optimization provides a more reliable solution for multi-output prediction under incomplete data conditions.

Despite these promising results, certain limitations warrant further investigation. In this study, TLBO was selected for hyperparameter optimization due to its simplicity and efficiency. The study focused on a single metaheuristic algorithm (TLBO), and future research could benefit from evaluating alternative optimization strategies. Other metaheuristic algorithms, such as Genetic Algorithm, Particle Swarm Optimization, or Differential Evolution, could be applied in future studies to assess potential differences in model performance. The proposed bias-corrected predictive imputation framework with metaheuristic-optimized boosting models offers a robust and accurate solution for multi-output prediction in the presence of missing values. By combining predictive imputation, bias correction, and metaheuristic optimization, this study contributes a comprehensive approach to mitigating data incompleteness and improving forecasting reliability, paving the way for future advancements in both theory and practice.

Future work may explore the extension of this framework to other domains and datasets, the comparison with additional metaheuristic algorithms, and the investigation of ensemble strategies that further enhance the stability of multi-output prediction under incomplete data conditions.

Acknowledgment

This research was partially supported by the Department of Computer, Faculty of Science and Technology and Research and Development Institute, Sakon Nakhon Rajabhat University, Sakon Nakhon, Thailand.

References

[1] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "Survey on Multi-Output Learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 7, pp. 2409–2429, 2020, doi: 10.1109/TNNLS.2019.2945133.

[2] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *J. Clin. Epidemiol.*, vol. 151, pp. 185–188, 2022, doi: https://doi.org/10.1016/j.jclinepi.2022.08.016.

[3] A. R. Alsaber, J. Pan, and A. Al-Hurban, "Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018)," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3. 2021. doi: 10.3390/ijerph18031333.

[4] H. Xue and Y. Niu, "Multi-Output Based Hybrid Integrated Models for Student Performance Prediction," *Applied Sciences*, vol. 13, no. 9. 2023. doi: 10.3390/app13095384.

[5] Z. Wang, Z. Feng, Z. Ma, and J. Peng, "A Multi-Output Regression Model for Energy Consumption Prediction Based on Optimized Multi-Kernel Learning: A Case Study of Tin Smelting Process," *Processes*, vol. 12, no. 1. 2024. doi: 10.3390/pr12010032.

[6] C. K. Enders, *Applied missing data analysis*. Guilford Publications, 2022.

[7] D. Adhikari *et al.*, "A Comprehensive Survey on Imputation of Missing Data in Internet of Things," *ACM Comput. Surv.*, vol. 55, no. 7, Dec. 2022, doi: 10.1145/3533381.

[8] A. D. Woods *et al.*, "Best practices for addressing missing data through multiple imputation," *Infant Child Dev.*, vol. 33, no. 1, p. e2407, 2024.

[9] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[10] R. Wu, S. D. Hamshaw, L. Yang, D. W. Kincaid, R. Etheridge, and A. Ghasemkhani, "Data Imputation for Multivariate Time Series Sensor Data With Large Gaps of Missing Data," *IEEE Sens. J.*, vol. 22, no. 11, pp. 10671–10683, 2022, doi: 10.1109/JSEN.2022.3166643.

[11] Y. Wang and E. Gao, "Research on Prediction of Missing Values Based on Multiple Models BT - Computer Applications," 2024, pp. 359–376.

[12] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[13] X. Zhang, C. Yan, C. Gao, B. Malin, and Y. Chen, “XGBoost imputation for time series data,” in *2019 IEEE International conference on healthcare informatics (ICHI)*, 2019, pp. 1–3.

[14] Z. E. Aydin and Z. K. Ozturk, “Performance analysis of XGBoost classifier with missing data,” *Manchester J. Artif. Intell. Appl. Sci.*, vol. 2, no. 02, p. 2021, 2021.

[15] G. Ke *et al.*, “Lightgbm: A highly efficient gradient boosting decision tree,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[16] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, “An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes,” *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 14, 2023.

[17] A. D. Hartanto, Y. N. Kholik, and Y. Pristyanto, “Stock price time series data forecasting using the light gradient boosting machine (LightGBM) model,” *JOIV Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2270–2279, 2023.

[18] R. V. Rao, V. J. Savsani, and D. P. Vakharia, “Teaching-learning-based optimization: A novel method for constrained mechanical design optimization problems,” *CAD Comput. Aided Des.*, vol. 43, no. 3, pp. 303–315, Mar. 2011, doi: 10.1016/j.cad.2010.12.015.

[19] M. Arashpour *et al.*, “Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization,” *Comput. Appl. Eng. Educ.*, vol. 31, no. 1, pp. 83–99, 2023.

[20] S.-H. Tseng and M.-H. Nguyen, “A metal price forecasting framework optimized with TLBO metaheuristic and selective for highest returns,” *Expert Syst. Appl.*, vol. 285, p. 127937, 2025, doi: https://doi.org/10.1016/j.eswa.2025.127937.

[21] T. Bui-Ngoc, D.-K. Ly, T. Nguyen, and T. Nguyen-Thoi, “Sustainable foundation design: Hybrid TLBO-XGB model with confidence interval enhanced load–displacement prediction for PGPN piles,” *Adv. Eng. Informatics*, vol. 65, p. 103288, 2025.

[22] E. Radmand, J. Pirgazi, and A. G. Sorkhi, “A Hybrid TLBO–XGBoost Model With Novel Labeling for Bitcoin Price Prediction,” *Int. J. Intell. Syst.*, vol. 2025, no. 1, p. 6674437, 2025.

[23] “Bias correction of numerical prediction model temperature forecast,” *UCI Machine Learning Repository*, 2020. https://doi.org/10.24432/C59K76