XGBoost optimization using hybrid Bayesian optimization and nested cross validation for calorie prediction

Budiman[1], Nur Alamsyah[2], Titan Parama Yoga[2], R. Yadi Rakhman Alamsyah[2], Elia Setiana[1]
[1]Department of Informatics Engineering, Faculty of Information and Technology, Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia
[2]Department of Information System, Faculty of Information and Technology, Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia

Article Info

Article history:

Received Aug 9, 2024
Revised Jan 23, 2025
Accepted Mar 23, 2025

Keywords:

Bayesian optimization
Calorie prediction
Extreme gradient boosting
Nested cross validation
Wearable devices data

ABSTRACT

Accurately predicting calorie expenditure is crucial for wearable device applications, enabling personalized fitness and health recommendations. However, traditional models struggle with high data variability and nonlinear relationships in activity data, leading to suboptimal predictions. This study addresses these challenges by integrating extreme gradient boosting (XGBoost) with Bayesian optimization and nested cross validation to enhance predictive accuracy. Unlike previous approaches, our method systematically tunes hyperparameters using Bayesian optimization while employing nested cross validation to prevent overfitting, ensuring robust model evaluation. We utilize a dataset of daily activity records, including steps, distance, and active minutes, extracted from wearable devices. Our experimental findings indicate a substantial enhancement in prediction performance, achieving a mean squared error (MSE) of 4294.27, an R-squared (R^2) score of 0.9917, and a root mean squared error (RMSE) of 65.53. The proposed model outperforms baseline approaches such as random forest and support vector machines in terms of predictive accuracy. These findings underscore the advantage of our approach in predictive modeling. Beyond calorie estimation, the proposed methodology is adaptable to other domains requiring high-precision predictions, such as healthcare analytics and personalized recommendation systems.

Corresponding Author:

Nur Alamsyah
Department of Information System, Faculty of Information and Technology
Universitas Informatika dan Bisnis Indonesia
St. Soekarno Hatta, Bandung, Indonesia
Email: nuralamsyah@unibi.ac.id

1. INTRODUCTION

The rapid advancement of wearable technology has transformed the health and fitness industry, enabling continuous real-time tracking of physical activity. Devices such as smartwatches and fitness trackers collect extensive data, including steps taken, distance traveled, active minutes, and calorie expenditure [1], [2]. Accurate calorie prediction is essential for personalized fitness recommendations, weight management, and preventive healthcare [3], [4]. Accurately predicting calorie expenditure is crucial for personalized fitness recommendations, weight manage- ment, and overall health improvement [5], [6]. Despite the accessibility of large amounts of data, developing robust models to predict daily caloric expenditure remains a challenging task due to the complex interactions of various factors that influence energy spending. A prominent issue in predicting calorie expenditure is the need for accurate and reliable models that can handle the high dimensionality and variability of the data [7]. Traditional methods, such as linear regression, often fail to capture the nonlinear relationships that are inher- ent in wearable data. More recent advancements in machine learning, particularly ensemble methods such as extreme gradient boosting (XGBoost), have shown promise in addressing this challenge [8]. However, the performance of these models relies heavily on precise hyperparameter tuning, which can be computationally expensive and vulnerable to overfitting [9], [10].

The literature on calorie prediction using machine learning has highlighted various approaches that have been applied. For instance, previous studies have used random forests and support vector machines, which have shown moderate success in calorie prediction [11]. One study by Hwang *et al.* [12] used random forests to predict calorie expenditure based on daily activity data, achieving an R-squared (R^2) score of 0.82. However, these methods often lack the robustness and predictive accuracy required for practical applications. In addition, another study by Lee [13] used linear regression and neural networks for calorie prediction, but faced similar challenges related to non-linear relationships and the complexity of data from wearable devices. This study showed that neural networks can provide better predictions than linear regression, but are still limited in handling high data variation [14]. Although XGBoost has been applied in various domains with significant success, its application in calorie prediction coupled with advanced hyperparameter optimization techniques is still rarely explored [15].

Research by Aziz *et al.* [16] showed that XGBoost has great potential in calorie prediction, achieving an R^2 score of 0.87 when applied to physical activity data. However, this study has not utilized more advanced hyperparameter optimization such as Bayesian optimization. These studies generally show that there is great potential in improving prediction accuracy through more advanced and integrated approaches. In this study, our main contribution is to propose a novel approach that leverages XGBoost with Bayesian optimization and nested cross validation to improve the accuracy of calorie expenditure prediction. Bayesian optimization systematically searches the hyperparameter space, finding optimal values more efficiently than traditional grid search methods. Nested cross validation further ensures that model performance is robustly validated, reducing the risk of overfitting and providing reliable estimates of model accuracy. To address these gaps, this study makes the following key contributions: i) utilization of XGBoost for calorie prediction, leveraging its ability to model nonlinear relationships effectively; ii) implementation of Bayesian optimization for hyperparameter tuning, systematically searching the hyperparameter space to enhance model efficiency; and iii) application of nested cross validation to improve model generalization and prevent overfitting, ensuring a robust evaluation framework.

Our proposed approach combines XGBoost with Bayesian optimization and nested cross validation, offering an innovative solution for high-accuracy calorie prediction from wearable data. By optimizing hyperparameters efficiently and incorporating a robust validation strategy, this study outperforms conventional methods and provides a more reliable predictive framework. The findings contribute to personalized health monitoring and can be extended to other domains requiring precise predictive modelling.

2. METHOD

The method used in this study is to predict calories based on daily activity data collected using Fitbit devices. The method consists of several main stages, namely data collection, data preprocessing, data splitting, model definition, hyperparameter tuning (using GridSearchCV and nested cross validation), and model evaluation. Our proposed method is presented in Figure 1.

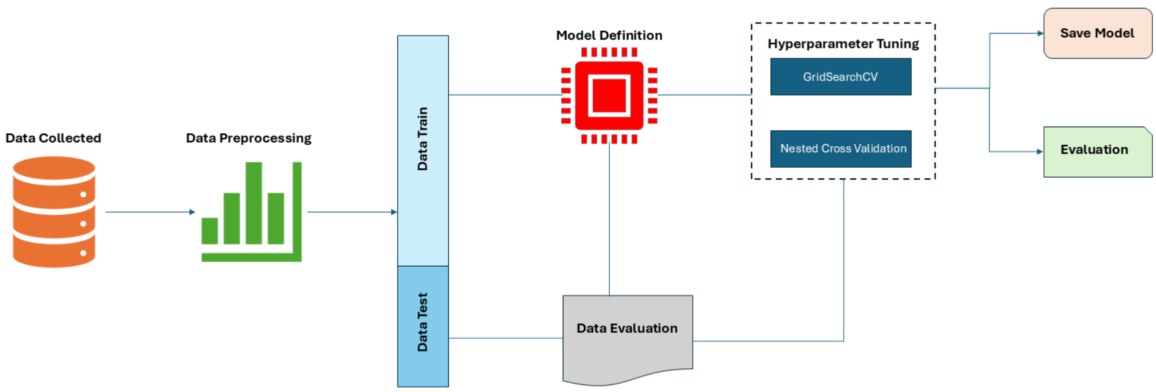

Figure 1. Proposed method

2.1. Data collected

The data used in this study was obtained from the Kaggle portal with the topic "Fitbit fitness tracker data: capstone project". This dataset contains users' daily activity records collected using Fitbit devices. The dataset consists of 940 rows and 15 columns that include various features such as number of steps, distance traveled, active minutes, and calories burned. This data provides comprehensive information about the user's daily physical activity, which is then used to predict calorie expenditure. A description of the features in the data used in this study is presented in Table 1.

Table 1. Feature descriptions for fitbit fitness tracker data

Variable	Description (reworded)
Id	A unique identifier assigned to each individual user.
ActivityDate	The specific date on which the recorded activity took place.
TotalSteps	The cumulative count of steps taken throughout the day.
TotalDistance	The total miles travelled within a single day.
TrackerDistance	The distance measured and logged by the tracking device.
LoggedActivitiesDistance	The distance covered during explicitly recorded activities.
VeryActiveDistance	The distance accumulated while performing highly vigorous activities.
ModeratelyActiveDistance	The total distance travelled during moderately intense movements.
LightActiveDistance	The measured distance from light-intensity activities.
SedentaryActiveDistance	The distance recorded while engaging in low-movement or stationary activities.
VeryActiveMinutes	The total time, in minutes, spent in high-intensity physical exertion.
FairlyActiveMinutes	The sum of minutes spent in moderately intense activities.
LightlyActiveMinutes	The duration of low-effort physical movement throughout the day.
SedentaryMinutes	The total amount of time spent in an inactive or resting state.
Calories	The total number of calories expended over the course of the day.

2.2. Data preprocessing

The data preprocessing stage is conducted to maintain data quality and ensure its readiness for model training [17]. The initial step involves addressing missing values, either by imputing them or discarding incomplete records. Subsequently, non-essential features, such as ActivityDate, are eliminated to prevent analytical inconsistencies. The dataset is then subjected to normalization or standardization, aligning all features to a uniform scale, which is particularly crucial for specific machine learning algorithms. Moreover, if categorical variables are present, they are transformed into numerical representations to facilitate model processing. The outcome of this preprocessing phase is a well-structured dataset, optimized for training and evaluation purposes.

2.3. Data splitting

After completing the data preprocessing stage, the dataset is then partitioned into two subsets: a training set and a testing set. This segmentation ensures that the model is evaluated on previously unseen data, allowing for a more reliable assessment of its performance [18]. In this study, the dataset was randomly split, with 80% allocated for training and 20% for testing, ensuring a well-balanced distribution. The training set is utilized to develop the model, while the testing set serves as an independent benchmark to measure its predictive accuracy.

2.4. Model definition

At this phase, the machine learning model to be used for the calorie prediction is defined. In this study, we used XGBoost, a powerful and efficient ensemble learning algorithm [19], [20]. XGBoost was chosen for its capabilities in handling complex data and producing accurate predictions with high computational speed [21]. The model is able to capture non-linear relationships in the data and automatically handle missing features, making it a suitable choice for calorie prediction based on physical activity data collected from Fitbit devices [22]. The basic parameters of the XGBoost model are set first, whichwill then be optimized in the next stage of hyperparameter tuning [23]. XGBoost is an ensemble-based machine learning algorithm that leverages decision trees as its foundational learning units [24]. It optimizes an objective function consisting of a loss function and a regulation term to prevent overfitting. The optimization function for XGBoost can be formulated as (1).

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

The function $(\mathcal{L}(\theta))$ represents the optimization objective, while $(l(\hat{y}_i, y_i))$ serves as the loss function, quantifying the discrepancy between the predicted value $(\hat{y}_i)$ and the actual value (y_i).

Additionally, $(\Omega(f_k))$ acts as the regularization component to mitigate overfitting. In regression tasks, one of the most commonly employed loss functions is the mean squared error (MSE), mathematically expressed as $(l(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2)$. The regularization term $(\Omega(f_k))$ is generally defined as (2).

$$(\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2) \tag{2}$$

(γ) and (λ) function as hyperparameters for regularization, while (T) denotes the total number of leaves in the decision tree. Additionally, (w_j) represents the weight assigned to the (j)-th leaf. Consequently, the complete optimization function for XGBoost in the context of regression with regularization can be expressed as (3).

$$(\mathcal{L}(\theta) = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \sum_{k=1}^{K}\left(\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2\right) \tag{3}$$

The XGBoost model aims to minimize this objective function during training to produce an optimal model. This approach allows XGBoost to handle complex data and provide accurate predictions with high computa- tional efficiency, making it a suitable choice for predicting calorie expenditure based on physical activity datacollected from Fitbit devices.

2.5. Hyperparameter tuning (GridSearchCV+nested cross validation)

To optimize the performance of the XGBoost model, hyperparameter tuning is carried out using GridSearchCV in combination with nested cross validation [25]. The purpose of hyperparameter tuning is to determine the best parameter set that minimizes the objective function, thereby enhancing both accuracy and robustness. GridSearchCV conducts a comprehensive search across a predefined parameter grid, training the model and assessing its performance through cross-validation for each parameter combination [26]. The most effective set of hyperparameters is then chosen based on the cross-validation results.

In this study, key hyperparameters that were fine-tuned include the learning rate (η), the maximum tree depth (max depth), and the number of trees (n estimators). To further improve model reliability and mitigate overfitting, nested cross validation is implemented. This method consists of two cross-validation loops: an outer loop dedicated to model evaluation and an inner loop for hyperparameter optimization. By employing this approach, an unbiased assessment of the model's performance is obtained. The optimization function for XGBoost, incorporating hyperparameter tuning, is expressed as (4).

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

Where the loss function $(l(\hat{y}_i, y_i))$ is defined as (5):

$$l(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2 \tag{5}$$

and the regularization term $(\Omega(f_k))$ is given by (6).

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

By minimizing this objective function, the hyperparameter tuning process aims to find the optimal values for (η), (max_depth), and (n_estimators), resulting in a model that generalizes well to new data.

2.6. Evaluation

After training and optimizing the XGBoost model using GridSearchCV and nested cross validation, the next phase involved assessing its performance. The evaluation was conducted using a pre-designated test dataset to verify the model's ability to accurately predict calorie expenditure on previously unseen data. Several essential metrics were utilized for performance assessment, including MSE, R^2 score, and root mean squared error (RMSE). These evaluation measures help determine the model's effectiveness in calorie prediction, ensuring high accuracy and strong generalization to new data. The results of the evaluation indicate that the proposed approach outperforms existing methods in calorie prediction based on physical activity data.

3. RESULTS AND DISCUSSION

The performance evaluation of the proposed model is conducted using various metrics, including MSE, R^2 score, and RMSE. These metrics offer a more thorough analysis of the model's precision and its capability to adapt to unseen data.

3.1. Result

The performance of different models was assessed using MSE, R^2 score, and RMSE. A summary of the evaluation results is presented in Figure 2. The random forest model, without hyperparameter tuning, produced an MSE of 69,109.31, an R^2 score of 0.86, and an RMSE of 262.89. This baseline model exhibited moderate performance but had potential for enhancement. The XGBoost model, optimized using GridSearchCV, demonstrated notable improvements, achieving an MSE of 50562.42, an R^2 score of 0.90, and an RMSE of 224.86. This indicates that XGBoost, even with basic hyperparameter tuning, outperforms the random forest model.

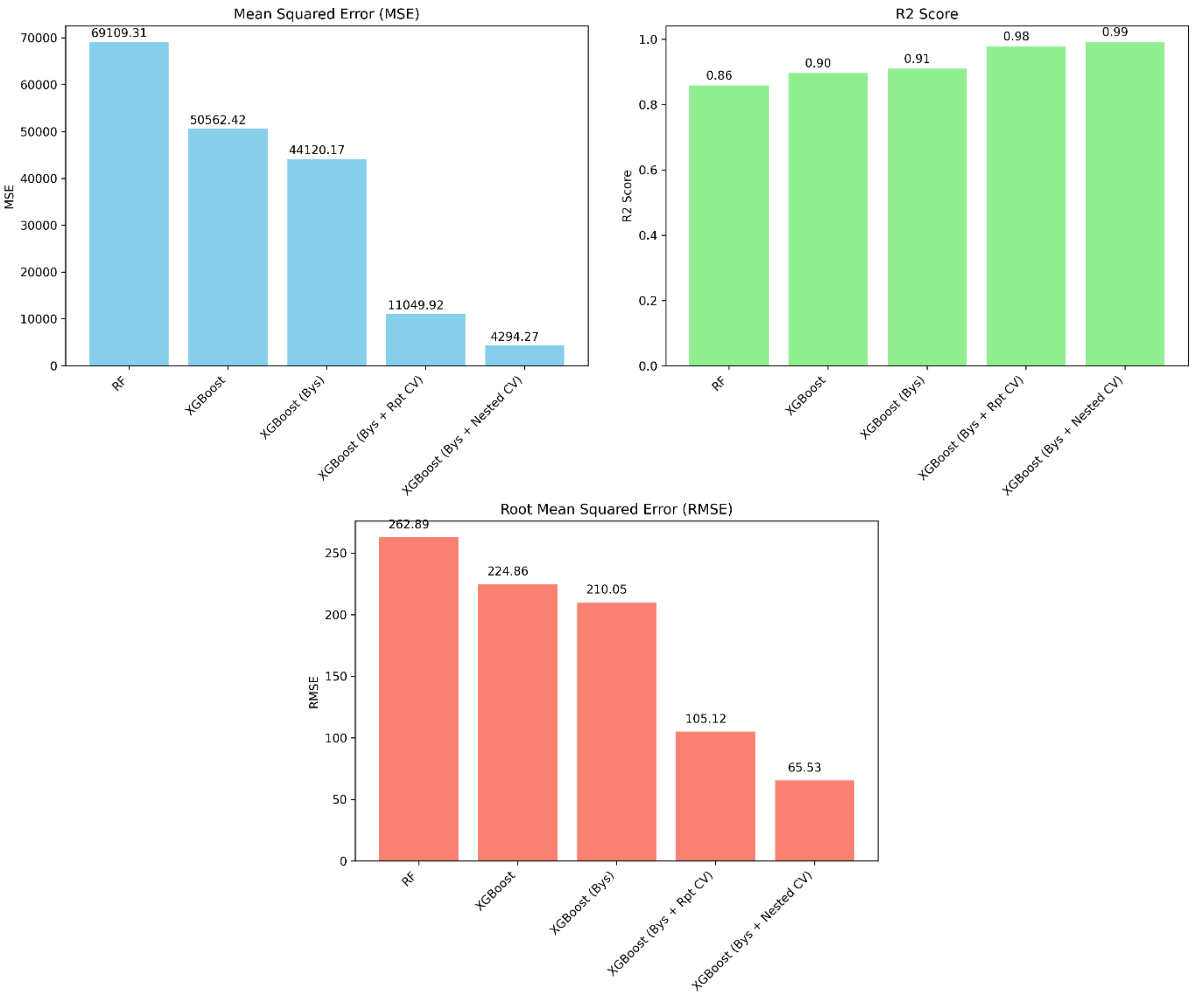

Figure 2. Evaluation result

Further enhancement was achieved with XGBoost combined with Bayesian optimization, resulting in an MSE of 44120.17, an R^2 score of 0.91, and an RMSE of 210.05. This demonstrates the effectiveness of Bayesian Optimization in improving the model's predictive accuracy. When repeated cross validation was added to the XGBoost model tuned with Bayesian optimization, the performance metrics improved sub-substantially, with an MSE of 11049.92, an R^2 score of 0.98, and an RMSE of 105.12. This approach shows a significant reduction in error and an increase in model reliability. The best results were obtained with XG-Boost using Bayesian optimization and nested cross validation. This model achieved an MSE of 4294.27, anR^2 score of 0.99, and an RMSE of 65.53. The combination of these advanced techniques provided the most accurate and robust model, highlighting their importance in model tuning and validation.

3.2. Discussion

The results of this study demonstrate significant improvements in predicting calorie expenditure using XGBoost with Bayesian optimization and nested cross validation. Our final model achieves superior performance, evidenced by the highest R^2 score of 0.99 and the lowest RMSE of 65.53. Comparing these results to similar studies highlights the advancements made in our ap- proach. For instance, Elshewey *et al.* [27] employed Random Forest and reported an R^2 score of 0.85 and an RMSE of 200. Their study, while comprehensive, did not integrate advanced hyperparameter tuning methods like Bayesian Optimization, which we found crucial for enhancing model performance. Another study by Wu [14] used support vector machines and achieved an R^2 score of 0.88. Although their model pro-vided reasonable predictions, it lacked the robustness and accuracy observed in our XGBoost model optimizedwith nested cross validation.

Further, a recent study by Asadi and Hajj [28] using neural networks achieved an R^2 score of 0.91 and an RMSE of 180. Their approach, while effective, underscores the potential of ensemble methods like XGBoost, especially when combined with sophisticated tuning techniques. Similarly, a study published in the International Research Journal of Modernization in Engineering Technology and Science (2023) highlighted the effectiveness of machine learning models, reporting an R^2 score of 0.93 using gradient boosting techniques, but without employing nested cross-validation for model validation, which might have resulted in a less reliableperformance estimation.

The novelty of our research lies in integrating Bayesian optimization with nested cross validation to develop an effective predictive model for calorie expenditure. Our experimental results indicate notable enhancements in prediction accuracy, achieving a MSE of 4294.27, an R^2 score of 0.9917, and RMSE of 65.53. These results highlight the potential of our approach in delivering more precise and dependable calorie predictions, which are essential for personalized fitness and health management.

4. CONCLUSION

This study successfully developed an advanced predictive model for calorie expenditure estimation using XGBoost with Bayesian optimization and nested cross validation. The proposed approach significantly improves prediction accuracy and robustness, as demonstrated by the final model achieving an R^2 score of 0.99 and an RMSE of 65.53, outperforming baseline models. These findings highlight the effectiveness of integrating hyperparameter tuning and robust validation techniques in predictive modeling. Beyond theoretical advancements, the proposed model has practical applications in real-time fitness tracking apps and health monitoring platforms. By providing more accurate calorie expenditure estimates, this model can enhance personalized fitness recommendations, support weight management programs, and assist healthcare providers in monitoring physical activity levels.

Despite its promising results, this study has certain limitations. First, the model relies on structured wearable data, which may not fully capture dynamic real-world variations in physical activity. Future research can explore deep learning approaches, such as long short-term memory (LSTM) or convolutional neural network (CNN) models, to handle more complex temporal patterns in calorie expenditure. Additionally, incorporating real-time data streams from internet of things (IoT)-enabled wearables could further improve prediction reliability. Another potential direction is applying this approach to other health metrics, such as heart rate variability, sleep tracking, or stress levels, expanding its utility beyond calorie estimation. Overall, the study demonstrates that advanced machine learning techniques can significantly enhance predictive accuracy in health and fitness applications. By leveraging continuously growing datasets from wearable technology, future research can refine these models to develop even more accurate, real-time, and personalized health monitoring systems.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Mrs. Elia Setiana from Universitas Informatika dan Bisnis Indonesia for his valuable assistance in data preprocessing and technical validation during the early phase of this research. The authors have obtained consent from all individuals mentioned in this section.

FUNDING INFORMATION

This research was funded by the Research and Community Service Division (PPM) of Universitas Informatika dan Bisnis Indonesia (UNIBI) under an internal grant of IDR 10,000,000.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Budiman	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Nur Alamsyah		✓				✓		✓	✓	✓	✓	✓		
Titan Parama Yoga	✓		✓	✓			✓			✓	✓		✓	✓
R. Yadi Rakhman Alamsyah	✓		✓	✓			✓			✓	✓		✓	✓
Elia Setiana					✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/arashnic/fitbit, reference title: Fitbit Fitness Tracker Data.

REFERENCES

[1] C. Del-Valle-Soto, R. A. Brisen˜o, L. J. Valdivia, and J. A. Nolazco-Flores, "Unveiling wearables: exploring the global landscape of biometric applications and vital signs and behavioral impact," *BioData Mining*, vol. 17, no. 1, p. 15, 2024, doi: 10.1186/s13040-024-00368-y.

[2] A. Fucarino *et al.*, "Emerging technologies and open-source platforms for remote physical exercise: Innovations and opportunities for healthy population—a narrative review," in *Healthcare*, vol. 12, no. 15, p. 1466, 2024, doi: 10.3390/healthcare12151466.

[3] G. Ji, J. Woo, G. Lee, C. Msigwa, D. Bernard, and J. Yun, "A IoT-Based Smart Healthcare in Everyday Lives: Data Collection and Standardization From Smartphones and Smartwatches," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 27597-27619, 2024, doi: 10.1109/JIOT.2024.3400509.

[4] K.-R. Hong, I.-W. Hwang, H.-J. Kim, S.-H. Yang, and J.-M. Lee, "Apple watch 6 vs. galaxy watch 4: A validity study of step-count estimation in daily activities," *Sensors*, vol. 24, no. 14, p. 4658, 2024, doi: 10.3390/s24144658.

[5] V. M. P. Cortes, A. Chatterjee, and D. Khovalyg, "Dynamic personalized human body energy expenditure: Prediction using time series forecasting lstm models," *Biomedical Signal Processing and Control*, vol. 87, 2024, doi: 10.1016/j.bspc.2023.105381.

[6] L.-I. Coman, M. Ianculescu, El.-A. Paraschiv, A. Alexandru, and I.-A. Bădărău, "Smart Solutions for Diet-Related Disease Management: Connected Care, Remote Health Monitoring Systems, and Integrated Insights for Advanced Evaluation," *Applied Sciences*, vol. 14, no. 6, p. 2351, 2024, doi: 10.3390/app14062351.

[7] M. Eed, A. A. Alhussan, A. T. Qenawy, A. M. Osman, A. M. Elshewey, and R. Arnous, "Potato consumption forecasting based on a hybrid stacked deep learning model," *Potato Research*, pp. 1–25, 2024, doi: 10.1007/s11540-024-09764-7.

[8] N. Alamsyah, Budiman, T. P. Yoga, and R. Y. R. Alamsyah, "A stacking ensemble model with smote for improved imbalanced classification on credit data," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 22, no. 3, pp. 657–664, 2024, doi: 10.12928/TELKOMNIKA.v22i3.25921.

[9] I. Priyana, N. Alamsyah, Budiman, A. P. Sarifiyono, and E. Rusnendar, "Predictive Boosting for Employee Retention with SMOTE and XGBoost Hyperparameter Tuning," *2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, Surakarta, Indonesia, 2024, pp. 92-97, doi: 10.1109/SIML61815.2024.10578116.

[10] M. Golyadkin, A. Gambashidze, I. Nurgaliev, and I. Makarov, "Refining the ONCE Benchmark with Hyperparameter Tuning," *IEEE Access*, vol. 12, pp. 3805-3814, 2024, doi: 10.1109/ACCESS.2023.3348750.

[11] N. U. Gilal *et al.*, "Evaluating machine learning technologies for food computing from a data set perspective," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 32041–32068, 2024, doi: 10.1007/s11042-023-16513-4.

[12] Y. -T. Hwang, Y. -R. Hsu, and B. -S. Lin, "Using B -Spline Model on Depth Camera Data to Predict Physical Activity Energy Expenditure of Different Levels of Human Exercise," in *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 79-88, Feb. 2024, doi: 10.1109/THMS.2023.3349030.

[13] K.-S. Lee, "Multi-spectral food classification and caloric estimation using predicted images," *Foods*, vol. 13, no. 4, p. 551, 2024, doi: 10.3390/foods13040551.

[14] H. Wu, "Intrusion Detection Model for Wireless Sensor Networks Based on FedAvg and XGBoost Algorithm," *International Journal of Distributed Sensor Networks*, no. 1, p. 5536615, 2024, doi: 10.1155/2024/5536615.

[15] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, "XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data," *2022 International Symposium on Electronics and Smart Devices (ISESD)*, Bandung, Indonesia, 2022, pp. 1-7, doi: 10.1109/ISESD56103.2022.9980630.

[16] M. T. Aziz, R. Sudheesh, R. D. C. Pecho, N. U. A. Khan, A. Ull, H. Era, and M. A. Chowdhury, "Calories burnt prediction using machine learning approach," *Current Integrative Engineering*, vol. 1, no. 1, pp. 29–36, 2023, doi: 10.59762/cie570390541120231031130323.

[17] N. Alamsyah, Saparudin, and A. P. Kurniati, "A Novel Airfare Dataset To Predict Travel Agent Profits Based On Dynamic Pricing," *2023 11th International Conference on Information and Communication Technology (ICoICT)*, Melaka, Malaysia, 2023, pp. 575-581, doi: 10.1109/ICoICT58202.2023.10262694.

[18] A. Alyaseen, A. Poddar, N. Kumar, P. Sihag, D. Lee, and T. Singh, "Assessing the compressive and splitting tensile strength of self-compacting recycled coarse aggregate concrete using machine learning and statistical techniques," *Materials Today Communications*, vol. 38, p. 107970, 2024, doi: 10.1016/j.mtcomm.2023.107970.

[19] Z. H. Wang, Y. F. Liu, T. Wang, J. G. Wang, Y. M. Liu, and Q. X. Huang, "Intelligent prediction model of mechanical properties of ultrathin niobium strips based on xgboost ensemble learning algorithm," *Computational Materials Science*, vol. 231, p. 112579, 2024, doi: 10.1016/j.commatsci.2023.112579.

[20] A. Bigdeli, A. Maghsoudi, and R. Ghezelbash, "A comparative study of the XGBoost ensemble learning and multilayer perceptron in mineral prospectivity modeling: a case study of the Torud-Chahshirin belt, NE Iran," *Earth Science Informatics*, vol. 17, no. 1, pp. 483–499, 2024, doi: 10.1007/s12145-023-01184-4.

[21] S. B. Jabeur, S. Mefteh-Wali, and J.-L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Annals of Operations Research*, vol. 334, no. 1, pp. 679–699, 2024, doi: 10.1007/s10479-021-04187-w.

[22] M. Y. Junior, R. Z. Freire, L. O. Seman, S. F. Stefenon, V. C. Mariani, and L. dos S. Coelho, "Optimized hybrid ensemble learning approaches applied to very short-term load forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 155, p. 109579, 2024, doi: 10.1016/j.ijepes.2023.109579.

[23] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, "An optimized xgboost-based machine learning method for predicting wave run-up on a sloping beach," *MethodsX*, vol. 10, p. 102119, 2023, doi: 10.1016/j.mex.2023.102119.

[24] S. Demir and E. K. Sahin, "An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost," *Neural Computing and Applications*, vol. 35, no. 4, pp. 3173–3190, 2023, doi: 10.1007/s00521-022-07856-4.

[25] T. O. Omotehinwa and D. O. Oyewola, "Hyperparameter optimization of ensemble models for spam email detection," *Applied Sciences*, vol. 13, no. 3, p. 1971, 2023, doi: 10.3390/app13031971.

[26] Y. Rimal and N. Sharma, "Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy," *Multimedia Tools and Applications*, vol. 83, no. 18, pp. 55091–55107, 2024, doi: 10.1007/s11042-023-17273-x.

[27] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, "Bayesian optimization with support vector machine model for Parkinson disease clas- sification," *Sensors*, vol. 23, no. 4, p. 2085, 2023, doi: 10.3390/s23042085.

[28] B. Asadi and R. Hajj, "Prediction of asphalt binder elastic recovery using tree-based ensemble bagging and boosting models," *Construction and Building Materials*, vol. 410, 2024, doi: 10.1016/j.conbuildmat.2023.134154.

BIOGRAPHIES OF AUTHORS

Budiman received his S.T. degree in Informatics Engineering from Pasundan University, in 2006, and his M.Kom. degree in Information Systems in 2013 from STMIK LIKMI, Indonesia. Currently, he is also a lecturer and researcher at Universitas Informatika dan Bisnis Indonesia at the Faculty of Technology and Informatics. His current research interests include machine learning, deep learning, data science, and system applications. He can be contacted at email: budiman@unibi.ac.id.

Nur Alamsyah received his S.T. degree in Informatics Engineering from Sekolah Tinggi Manajemen Informatika Bandung, in 2002, and his M.Kom. degree in Information Systems in 2013 from Sekolah Tinggi Manajemen Informatika LIKMI Bandung, Indonesia. He is currently continuing his Doctor of Informatics Study at Telkom University Bandung. He is also a lecturer and researcher at Universitas Informatika dan Bisnis Indonesia at the Faculty of Technology and Informatics. His current research interests include data science, machine learning, deep learning, optimization and text mining. He can be contacted at email: nuralamsyah@unibi.ac.id.

Titan Parama Yoga received his S.Kom. degree in accounting computerization (specialization in informatics management) from STMIK Bandung, in 2001 and his M.Kom. degree in information systems in 2016 from STMIK LIKMI. Currently he is also a lecturer and researcher at the University of Informatics and Business Indonesia at the Faculty of Technology and Informatics. His current research interests include information systems audit, UI/UX, and data science. He can be contacted at email: titanparamayoga@gmail.com.

R. Yadi Rakhman Alamsyah received his S.T. degree in Informatics Engineering from STMIK Bandung, in 2007, and his M.Kom. degree in Information Systems in 2017 from STMIK LIKMI, Indonesia. Currently, he is also a lecturer and researcher at Universitas Informatika dan Bisnis Indonesia at the Faculty of Technology and Informatics. His current research interests include software engineering, algorithm data, and multimedia. He can be contacted at email: r.yadi@unibi.ac.id.

Elia Setiana received her S.Kom. degree in Informatics from Langlangbuana University Bandung in 2000 and her M.T. degree in the same field in 2015. Currently, she is also a lecturer and researcher at Universitas Informatika dan Bisnis Indonesia in the Faculty of Technology and Informatics. His current research interests include data science and software engineering. He can be contacted at email: elia.setiana@unibi.ac.id.