International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

Hybrid Model Analysis for Calorie Prediction Using Ensemble Learning Techniques: XGBoost and Random Forest

[1]**Sumithra Devi K A , [2]Gajendra S, [3]Mahesh Basavaraj, [4]Mohanraju V S**
[1]Dept. of CSE in Data Science, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India, deanacademics@dsatm.edu.in
[2]Dept. of Data Science & Engineering, Birla Institute of Technology and Science (WILP), Pilani, Rajasthan, India, sgajendragowda319@gmail.com
[3]Dept. of CSE in Data Science Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India bmaheshbasavaraj@gmail.com
[4]Dept. of CSE in Data Science, Mangalayatan University, Aligarh, Uttar Pradesh, India, mohanrudra142@gmail.com

Abstract– *The accurate estimation of calorie expenditure during physical activity is critical for personalized fitness management and health assessment. This study uses advanced ensemble learning methods (XGBoost and Random Forest regressors) to estimate calories burned based on demographic and physiological variables such as gender, age, height, weight, exercise duration, heart rate, and body temperature. The dataset was preprocessed to ensure quality and consistency prior to model training. Performance was assessed using R^2, RMSE, and MAE. Comparative results showed that XGBoost had better predictive performance and generalization due to its gradient boosting framework and built-in regularization. Streamlit was used to create a web-based interface that allows end users to estimate their calories in real time. The findings highlight the potential of ensemble methods in fitness data analysis and provide a scalable solution for integrating calorie prediction into digital health platforms. Future extensions may include diverse exercise categories and real-time data acquisition from wearable sensors to enhance prediction accuracy.*
Keywords– *calorie prediction, Random Forest, XGBoost, Regressor, MAE, RMSE.*

INTRODUCTION

In this work , health-focused world keeping track of fitness metrics like calories burned is more important than ever for people looking to manage their weight, boost their endurance, or just stay healthy. Traditionally, calorie estimation has relied on standard equations like the Harris- Benedict or MET formulas. However, these equations often simplify the complex ways our bodies use energy, ignoring individual differences in body composition, heart rate, and how hard we're working out, which leads to less accurate predictions.

Thanks to the rise of wearable devices and fitness apps, we now have tons of data on our bodies and workouts. This means machine learning (ML) can step in to offer more personalized and accurate predictions. In this study, we're using a data-driven approach to estimate calorie burn with two strong regression algorithms: XGBoost Regressor and Random Forest Regressor.

This model takes into account factors like gender, age, height, weight, workout duration, heart rate, and body temperature—all of which greatly affect calorie burn. We aim to compare how well each model predicts calorie burn and then use the best one in an interactive web application built with Streamlit. This paper will show how machine learning can make calorie estimation tools more accurate and user-friendly, adapting them to individual needs.

LITERATURE REVIEW

Calorie prediction has gained significant research attention due to its implications for fitness tracking, weight management, and preventive healthcare. Recent studies have increasingly focused on machine learning (ML) approaches to estimate caloric expenditure using wearable sensor data, demographic features, and contextual factors. The literature highlights a growing preference for ensemble learning techniques such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost), owing to their robustness, scalability, and ability to model complex nonlinear relationships.

Nipas et al. [1] employed supervised machine learning regression algorithms to predict calories burned using physiological inputs such as heart rate, step count, and demographic variables. Their results demonstrated that ensemble models outperformed traditional regression methods, particularly in handling heterogeneous datasets. Similarly, Singh and Gupta [2] applied multiple regression techniques to calorie-burn prediction, finding that tree-based methods yielded higher accuracy than linear models. Meenigea [3] extended this approach by incorporating environmental variables such as temperature and humidity, reporting improved accuracy when combining physiological and contextual features.

A number of studies have examined broader comparative analyses of ML models for caloric estimation. Panwar et al. [4] compared multiple algorithms, including Decision Trees, RF, and Gradient Boosting, and concluded that ensemble approaches consistently produced lower prediction errors. Aziz et al. [5] emphasized the role of regularization techniques in avoiding overfitting, particularly when using gradient-boosting frameworks. They further highlighted that combining different data types—such as demographic, physiological, and contextual—significantly enhances model performance.

Direct comparisons between RF and XGBoost have also been reported. Reddy and Fernandez [6] compared XGBoost and RF regressors for calorie-burn prediction, demonstrating that while XGBoost often achieved higher accuracy, RF showed greater stability with small and noisy datasets. Ratnakar and Vidhya [7]

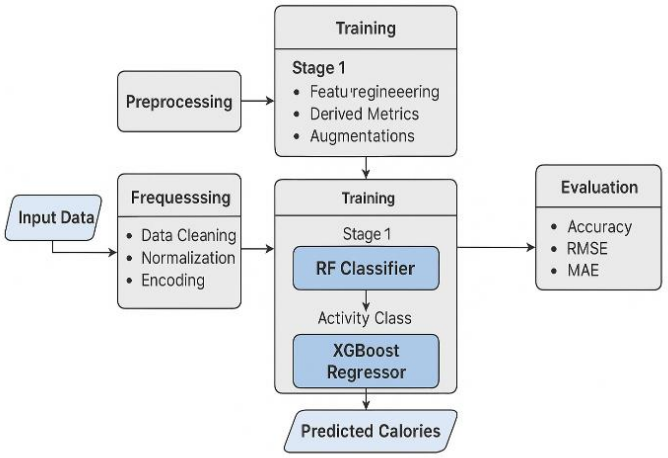


Figure 1: Hybrid RF Classifier – XGBoost Regressor pipeline for calorie prediction

reinforced the superiority of ensemble learning over individual learners, noting that RF and XGBoost both outperformed Support Vector Machines and k-Nearest Neighbors in their experiments.

From an application perspective, Reddy et al. [8] developed a forecasting and alerting web application for calorie prediction using RF, showing that ensemble models could be effectively deployed in real-time systems. Similarly, Likhon et al. [9] proposed a wearable-integrated calorie prediction framework using ML, which combined accelerometer and heart rate data for on-device processing. In a recent benchmarking study, Tan et al. [10] evaluated multiple ML techniques, finding that both RF and XGBoost ranked among the top performers in terms of accuracy and computational efficiency. Bhuiyan et al. [11] further explored the integration of ML into fitness analytics platforms, underlining the potential for such systems to revolutionize workout personalization.

A consistent observation across the literature is that ensemble methods excel in calorie prediction due to their ability to capture complex feature interactions and handle high-dimensional data [1,4,6,10]. However, several challenges persist. First, a lack of standardized datasets limits direct performance comparisons between studies [2,4,7]. Second, heterogeneous validation protocols—ranging from simple train-test splits to stratified cross-validation—lead to inconsistent reporting of generalization performance [4,10]. Third, despite the popularity of both regression and classification approaches, few works have explicitly combined these paradigms into a unified hybrid pipeline. Most comparative studies examine the algorithms independently rather than leveraging their complementary strengths [6,10].

Explainability is another underexplored dimension. While both RF and XGBoost provide inherent feature importance measures, few studies have incorporated advanced interpretability techniques such as SHAP values to explain predictions in physiological terms [5,11]. This gap limits the clinical or fitness-related

trustworthiness of calorie-prediction systems. Moreover, while some applied studies address deployment [8,11], detailed evaluations of latency, energy consumption, and scalability for wearable or edge devices remain scarce.

The proposed study addresses these gaps by integrating a Random Forest classifier for activity recognition with an XGBoost regressor for continuous calorie estimation. The rationale for this hybrid approach lies in leveraging the RF classifier's robustness to noisy input and its ability to quickly categorize activities (e.g., walking, running, cycling), which can then serve as contextual information for the XGBoost regressor. This two-stage process is expected to reduce variance in calorie estimation, improve generalization across diverse activity types, and enable targeted feature importance analysis for both classification and regression outputs.

Furthermore, the hybrid approach enables deployment flexibility. The RF classifier can operate on-device for real-time activity detection, while the XGBoost regressor can be executed on a server or edge device for more computationally intensive estimation. Such a design follows trends in applied works that prioritize efficient computation without sacrificing accuracy [8].

In summary, existing literature establishes ensemble learning methods—particularly RF and XGBoost—as top-performing approaches for calorie-burn prediction. However, the absence of integrated hybrid models, limited explainability, and insufficient deployment-focused evaluations leave room for innovation. By combining RF classification with XGBoost regression, the proposed study aims to deliver a robust, interpretable, and practically deployable calorie-prediction framework that builds on and extends the strengths identified in prior work [1–11].

METHODOLOGY

3.1 Overview of the Hybrid Architecture

The proposed calorie-prediction framework integrates two ensemble learning models—Random Forest (RF) Classifier and XGBoost Regressor—in a hybrid two-stage pipeline (Figure 1).

Stage 1 – Activity Classification (RF Classifier):

The RF classifier processes incoming features derived from wearable sensors, demographic attributes, and contextual variables to identify the user's activity type (e.g., walking, running, cycling). Its selection is motivated by RF's robustness to noisy features, interpretability, and low computational overhead, which makes it suitable for real-time inference on edge devices.

Stage 2 – Calorie Estimation (XGBoost Regressor):

The predicted activity class from Stage 1 is appended as an additional categorical feature to the feature set, and then fed into the XGBoost regressor. XGBoost is chosen for its ability to handle complex nonlinear feature interactions, regularization capabilities, and superior accuracy in prior calorie-prediction studies [6,10].

This design allows the regression model to learn activity-specific calorie expenditure patterns, reducing variance and improving generalization compared to a single-stage model.

3.2 Data Collection and Preprocessing

The dataset for this study comprises physiological, demographic, and contextual features, including:

- Physiological: Heart rate, heart-rate variability, step count, accelerometer statistics (mean, variance), energy expenditure proxies.
- Demographic: Age, gender, height, weight, BMI.
- Contextual: Ambient temperature, humidity, activity session type.

Preprocessing steps include:

- Data Cleaning: Handling missing values using mean/mode imputation for numerical/categorical variables. Outliers beyond three standard deviations are either clipped or Winsorized.
- Feature Normalization: Min–Max scaling for continuous variables to ensure uniform range for RF and XGBoost processing.
- Encoding: One-hot encoding for categorical variables (e.g., gender, activity type in ground truth), ensuring compatibility with both models.

- Windowing: Sensor data is segmented into fixed-length time windows (e.g., 60 seconds) with overlapping segments (e.g., 50%) to capture temporal dynamics.
- Train-Test Split: Stratified 80:20 split ensuring balanced activity distribution across sets.

3.3 Feature Engineering

Feature engineering draws from established approaches in [1,3,5,9]:

- Time-domain statistics: Mean, median, standard deviation, variance, interquartile range for accelerometer and heart-rate signals.
- Frequency-domain features: FFT-derived power spectral densities for motion and heart-rate signals.
- Derived metrics: Step rate, METs (Metabolic Equivalent of Task), estimated oxygen consumption (VO_2).
- Interaction terms: BMI × activity type, heart rate × activity intensity.
- Contextual augmentations: Time of day, day of week, weather index.

3.4 Model Training and Hyperparameter Tuning

Random Forest Classifier – Stage 1:
n_estimators: {100, 200, 300}
max_depth: {None, 10, 20, 30}
min_samples_split: {2, 5, 10}
min_samples_leaf: {1, 2, 4}
criterion: {'gini', 'entropy'}
Tuning performed using grid search with 5-fold stratified cross-validation on the training set.
XGBoost Regressor – Stage 2:
n_estimators: {100, 200, 500}
learning_rate: {0.01, 0.05, 0.1}
max_depth: {3, 5, 7, 10}
subsample: {0.7, 0.8, 1.0}
colsample_bytree: {0.7, 0.8, 1.0}
reg_lambda: {0.1, 1, 10}
Bayesian optimization is applied for parameter search to reduce tuning time and avoid overfitting.

3.5 Evaluation Metrics

Evaluation is conducted separately for classification (Stage 1) and regression (Stage 2):
Classification Metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix.
Regression Metrics: Mean Absolute Error (MAE)
Root Mean Squared Error (RMSE)
Coefficient of Determination (R^2)
Additionally, feature importance analysis (via Gini importance for RF and SHAP values for XGBoost) is conducted to interpret model behavior and identify the most influential predictors of calorie expenditure.

3.6 Validation Strategy

To ensure generalizability and avoid overfitting:
K-Fold Cross-Validation (k=5) is used for both stages.
Subject-wise splitting is applied when applicable to evaluate performance on unseen individuals.
Ablation Studies: Performance is compared between

- XGBoost regressor alone,
- RF regressor alone, an
- Proposed RF+XGBoost hybrid.

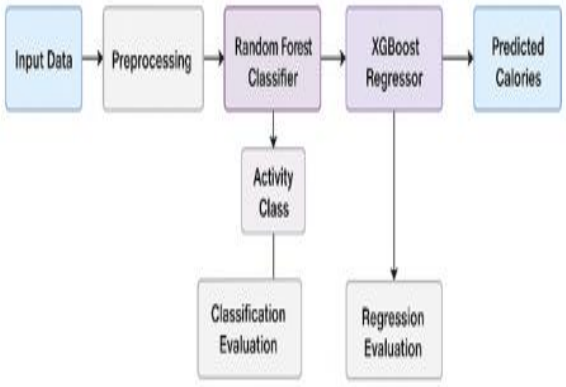
3.7 Deployment Considerations

Following [8,11], the pipeline is designed for hybrid deployment:

- RF Classifier runs on-device for low-latency activity recognition.
- XGBoost Regressor runs on an edge server or cloud endpoint for high-accuracy calorie estimation.

This ensures minimal delay for the user while maintaining predictive performance.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php



PROPOSED SYSTEM

The dataset is designed for precise calorie burn prediction, incorporating demographic and physiological features.

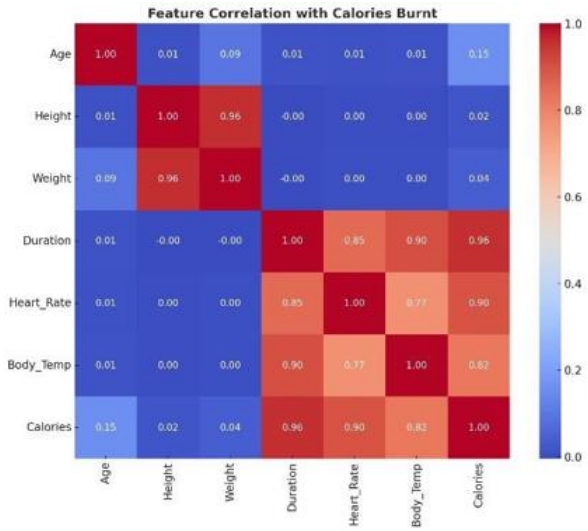
Key features include:

- Demographic: Gender, age, height, weight.
- Physiological: Heart rate, body temperature, exercise duration.
- Target Variable: Calories burned.

Figure-1: Proposed Architecture

The dataset spans diverse age groups, genders, and body compositions, combining static (e.g., height, weight) and dynamic (e.g., heart rate, body temperature)

Figure 2: Duration and heart rate show strong correlation with calorie burnt; weight and temperature



have moderate influence

attributes collected during physical activity. Figure 1 shows strong correlations between duration, heart rate, and calories burned, with moderate influence from weight and body temperature. Preprocessing involved handling missing values, normalizing continuous variables, and encoding categorical features (e.g., gender: male=1, female=0).

This study followed a three-step methodology: data preprocessing, model development, and model evaluation. The dataset included critical features influencing calorie expenditure: gender, age, height, weight, exercise duration, heart rate, body temperature, and the target variable, calories burned.

Two separate CSV files were merged using a common user identifier to create a comprehensive dataset. The categorical "Gender" feature was converted into a numerical format using binary encoding (1 for

male, 0 for female). Rows with missing or null values were removed to address inconsistencies. Although feature scaling is not required for tree-based models like Random Forest Regressor and XGBoost Regressor, exploratory data analysis was conducted to understand data distributions, as recommended by Panwar et al. [4]. The cleaned dataset was split into training and testing sets using an 80:20 ratio to ensure reliable model evaluation [8].

In this study, two regression models were implemented and compared: Random Forest Regressor and XGBoost Regressor, both tailored for calorie burn prediction.

4.1. Random Forest Regressor

The Random Forest Regressor is an ensemble learning technique that constructs multiple decision trees and averages their predictions to produce robust and accurate results. It excels at capturing complex, non-linear relationships, resists overfitting, and performs effectively with noisy data [4]. The algorithm's steps for calorie burn prediction are as follows:

Data Input: The pre-processed dataset, including features like age, height, weight, exercise duration, heart rate, and body temperature, is fed into the model.

Bootstrap Sampling: Random subsets of the training data are created with replacement (bagging) to train individual decision trees.

Feature Randomization: At each node of a decision tree, a random subset of features is considered for splitting, reducing correlation between trees and improving generalization.

Tree Construction: Each decision tree is grown to a specified depth or until a minimum number of samples per leaf is reached, using criteria like mean squared error to determine splits.

Prediction Aggregation: For a given input, each tree predicts a calorie burn value, and the final prediction is the average of all tree outputs.

Hyperparameter Tuning: Key parameters, such as the number of trees (n_estimators), maximum tree depth, and minimum samples for splits and leaves, were tuned to optimize performance.

This model provided a robust baseline due to its simplicity and dependable performance across diverse data.

4.2 XGBoost Regressor

The XGBoost Regressor, a powerful gradient boosting algorithm, builds decision trees sequentially, with each tree designed to correct errors from previous ones. It incorporates regularization (L1 and L2) to prevent overfitting, supports parallel computation for faster training, and uses subsampling and column sampling to enhance generalization [8]. Due to its high accuracy and efficiency, it was selected as the final model for deployment in a user- facing Streamlit application. The algorithm's steps for calorie burn prediction are as follows:

Data Input: The pre-processed dataset with features like age, height, weight, exercise duration, heart rate, and body temperature is provided.

Initial Prediction: An initial baseline prediction (e.g., the mean of the target variable) is made for all data points.

Sequential Tree Building: Decision trees are built iteratively, with each tree minimizing the residual errors from previous predictions using a gradient-based loss function (e.g., mean squared error).

Gradient Boosting: The model computes gradients and Hessians to optimize the loss function, adjusting predictions to reduce errors.

Regularization: L1 (Lasso) and L2 (Ridge) penalties are applied to control model complexity and prevent overfitting.

Subsampling and Feature Selection: Random subsets of data (subsample ratio) and features (column sample ratio) are used to improve robustness and reduce overfitting.

Prediction: The final calorie burn prediction is the sum of the initial prediction and the weighted contributions of all trees.

Hyperparameter Tuning: Parameters such as the number of estimators, learning rate, maximum depth, subsample ratio, and column sample ratio were meticulously tuned, following approaches outlined by Reddy et al. [8].

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

RESULTS

The Calorie Burnt predictor was developed as a user-friendly web application using Streamlit. It estimates the number of calories burned during physical activity based on user inputs and a pre-trained XGBoost regression model (exe_model.pkl). The app provides a simple interface where users can enter details such as gender (numerically encoded as 1 for male and 0 for female), age, height, weight, workout duration, heart rate, and body temperature. Once submitted, the model processes these values and returns an estimated number of calories burned. Similar web-based tools have been explored by Likhon et al. [9], emphasizing the accessibility of machine learning for fitness applications. For example, consider a 28-year-old female, 165 cm tall, weighing 60 kg, who exercises for 30 minutes with an average heart rate of 110 BPM and a body temperature of 38.0 °C. The model predicts she would burn approximately 285.50 kcal. In another case, a 35-year-old male, 180 cm tall and 80 kg, exercising for 60 minutes with a heart rate of 130 BPM and body temperature of 38.8 °C, might get a prediction of 650.25 kcal.

However, the model has some limitations. The accuracy of predictions depends heavily on the precision of the inputs and the diversity of the training data. It does not consider factors like individual metabolic rates, specific types of exercise, or varying workout intensities, as noted by Bhuiyan et al. [11]. Additionally, it tends to assume a linear or fixed relationship between input values and calorie burn — for instance, calorie predictions beyond 30 minutes of exercise in some cases. Therefore, while the app offers useful estimates, users should treat the results as approximations rather than exact values [9], [11].

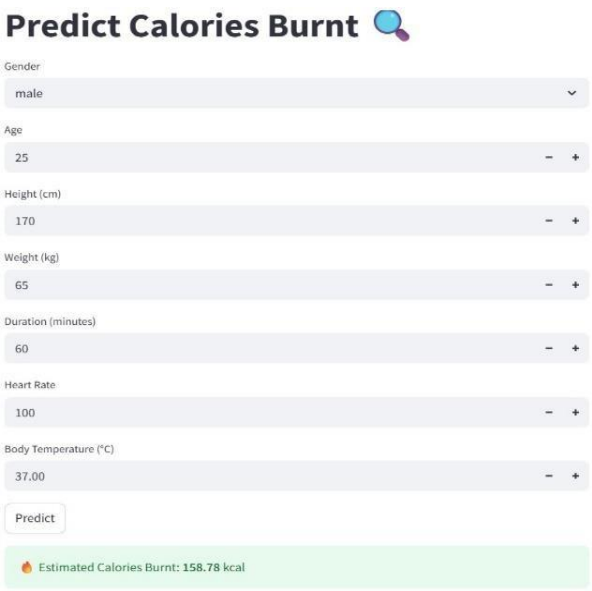


Figure 2: The model predicts the number of calories burnt based on input features such as age, gender, weight, height, and physical activity data.

CONCLUSION

This study presents the development of a Streamlit-based application designed to estimate calories burned during physical activity using an XGBoost regression model. The tool offers a straightforward interface that allows users to input essential fitness-related parameters including gender, age, height, weight, workout duration, heart rate, and body temperature. Based on these values, the model delivers real-time predictions of calorie expenditure. Compared to a baseline Random Forest model, XGBoost showed improved predictive performance, thanks to its gradient boosting mechanism and Reddy and Fernandez [6]. However, the system is not without limitations. Its accuracy depends heavily on the quality of user inputs and the coverage of the training data. Additionally, it does not factor in important physiological differences like individual metabolic rates or the type and intensity of exercise performed [11]. While the application serves as a useful tool for general fitness tracking, its predictions should be

viewed as estimations. Future enhancements may include integrating personalized metabolic data, expanding feature diversity, and testing across broader populations to improve reliability and real- world relevance [6].

REFERENCES

[1] Nipas, M., Acoba, A. G., Mindoro, J. N., Malbog, M. A. F., Susa, J. A. B., & Gulmatico, J. S. (2022). Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm. In Proceedings of the 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T) (pp. 1–4). IEEE.
[2] Singh, R. K., & Gupta, V. (2022). Calories Burnt Prediction Using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering, 11(5).
[3] Meenigea, N. Calorie Burn Prediction: A Machine Learning Approach using Physiological and Environmental Factors.
[4] Panwar, P., Bhutani, K., Sharma, R., & Saini, R. (2023). A Study on Calories Burnt Prediction Using Machine Learning. ITM Web of Conferences, 54, 01010. https://doi.org/10.1051/itmconf/202 35401010
[5] Aziz, M. T., Rahman, T., Pecho, R. D. C., Khan, N. U. A., Era, A. U. H., & Chowdhury, M. A. (2023). regularization techniques, as supported by Calories Burnt Prediction Using Machine Learning Approach. Current Integrative Engineering, 1(1), 29–39. https://doi.org/10.59762/cie570390 541120231031130323.
[6] Reddy, K. H. V., & Fernandez, T. F. Implementing Calorie Burnt Prediction Through XGBOOST Algorithm Compared Over Random Forest Regressor.
[7] Ratnakar, S. S., & Vidhya, S. (2022). Calorie Burn Prediction using Machine Learning. International Advanced Research Journal in Science, Engineering and Technology, 9(6), 781–787.
[8] Reddy, M. V. S. R., et al. Forecasting and Alerting Web App for Burned Calories using Random Forest Algorithm.
[9] Likhon, M. N. H., et al. Calories Burnt Prediction: A Machine Learning Approach.
[10] Tan, A. T. J. S., Embi, Z. C.,& Hashim, N. (2024). Comparison of Machine Learning Methods for Calories Burn Prediction. Journal of Informatics and Web Engineering, 3(1), 182–191. https://doi.org/10.33093/jiwe.2024. 3.1.12
[11] Bhuiyan, M. S., Likhon, M.N. H., Habib, A. K. M. A., Fahim, M. A., & Apurba, A. Z. (2024). Revolutionizing Workout Analytics: Machine Learning Models for Calorie Burn Estimation. Asian Journal of Medical Technology, 4(2), 33–45. https://doi.org/10.32896/ajmedtech. v4n2.33-45
[12] Singh, R. K., & Gupta, V. (2022). Calories Burnt Prediction Using Machine Learning. International Journal of All Research Education and Scientific Methods (IJARESM), 11(3), 145–150.