Machine Learning-Based Calorie Burn Prediction for Gym Members

GSS Senevirathne
Department of Electronic, Electrical and Telecommunication Engineering
General Sir John Kotelawala Defence University
Ratmalana, Sri Lanka.
39-eng-0002@kdu.ac.lk

***Abstract*—This study focuses on predicting the calories burned by gym members during their exercise sessions using machine learning models. Knowing how many calories are burned helps gym members track their progress, adjust their workouts, and stay motivated toward their fitness goals. In this study, three different models were built and tested: Random Forest, Gradient Boosting, and Linear Regression. Each model's performance was measured using metrics like R^2, RMSE, and MAE. After tuning the model parameters for better accuracy, the Gradient Boosting model emerged as the best performer, with the highest R^2 score and the lowest error rates. The process included data preparation, splitting the data for training and testing, applying cross-validation, and fine-tuning hyperparameters to improve results. Overall, this work shows that machine learning, especially Gradient Boosting, can effectively predict burning calories, providing gym members with useful insights to enhance their training routines.**
Keywords—Machine learning, Gradient Boosting Regressor, Regression, Calories burn prediction

INTRODUCTION

Balancing the calories we consume with the calories we burn is key to maintaining a healthy body and achieving fitness goals. For people who exercise regularly, especially gym members, understanding how much energy they use during exercise is important for tracking progress, adjusting exercise routines, and managing their diet. Calories burned during a workout are a direct measure of energy expenditure, and knowing this information can help people make better decisions about their health.[1] Many gym users expect that structured workouts such as cardio, strength training, high-intensity interval training (HIIT), or yoga burn more calories than regular physical activity, such as walking or housework. While this is often true, the actual number of calories burned can vary significantly from person to person and from exercise to exercise.

Several factors affect how many calories are burned during exercise. Without a reliable way to measure or estimate calorie burn, gym members can struggle to understand whether their workouts are truly effective or how to adjust their routines to meet their goals. To address this issue, this study focuses on developing a machine learning model that can predict the number of calories burned during an exercise session based on a person's physical and exercise-related data. The dataset used in this project was obtained from Kaggle and contains 973 records [4]. Each entry represents a single exercise session and includes ten key features: gender, age, height, weight, exercise duration, average heart rate (BPM), maximum heart rate, resting heart rate, exercise type (cardio, strength, HIIT, or yoga), and experience level (rated from 1 to 3)[3].The target variable is the number of calories burned during that session. These features were chosen because they are known to have a strong relationship with energy expenditure. In this study, we also examine how each of these features relates to calorie burn, helping us better understand their individual and combined effects on energy expenditure.

Three machine learning models were selected for comparison in this study: linear regression, random forest regression, and Gradient boosting regression. These models are trained and tested on the dataset to determine which one provides the most accurate predictions. The goal is to create a practical tool that gym members can use to estimate how many calories they are likely to burn during a session based on their personal data and workout details. This type of tool can help users plan their fitness routines more effectively, track progress with better accuracy, and make diet and training adjustments as needed.

PROBLEM SELECTED

A clear understanding of energy expenditure during exercise is essential to achieving fitness and weight management goals. Gym members often rely on average estimates or fitness trackers to measure calories burned, but these tools may not consider individual differences such as age, weight, heart rate, exercise type, and duration. As a result, the feedback they provide can be inaccurate or misleading, affecting exercise planning and food choices. This study addresses the problem of inaccurate calorie estimates by developing a machine learning model that predicts calories burned during exercise using personal and session-specific data. The goal is to provide exercisers with a more accurate and personalized tool to monitor their energy use and improve fitness outcomes.

TYPE OF MACHINE LEARNING SOLUTION

This project uses a supervised machine learning approach because it relies on a labeled dataset that already includes both input features (such as age, weight, heart rate, and exercise type) and the corresponding target output (calories burned). Since the goal is to predict a continuous numerical value, calorie burn, the problem falls under regression, rather than classification.

Three regression models were selected for this study: linear regression, random forest regression, and sequential boosting regression, each of which serves a unique purpose in the modeling process. Linear regression is used as a base model due to its simplicity, fast training time, and high interpretability. It works best when the relationship between variables is linear and does not require high-level parameter tuning, making it a useful point of comparison.

Random Forest Regression is chosen because it handles mixed data types naturally, captures nonlinear relationships, and is less prone to overfitting. It also provides feature importance, which helps to understand which factors contribute the most to calorie burn. It works well on medium-sized datasets, such as the one used in this project.[2]

Gradient Boosting Regression is included because it often provides the highest prediction accuracy. It builds models sequentially, learns from previous errors, and can capture complex, nonlinear interactions between features. However, it requires careful hyperparameter tuning to avoid overfitting.

TRAINING AND TESTING OF THE MODEL

To train and test the machine learning models, several Python libraries were used. These include pandas for data handling, NumPy for numerical operations, matplotlib and seaborn for data visualization, and scikit-learn for implementing and evaluating machine learning algorithms. The dataset was obtained from Kaggle and contains workout session data from 973 gym members. A total of 11 features were selected for analysis: gender, age, height, weight, session duration (in hours), average BPM (beats per minute during the workout), maximum BPM, resting BPM (before the workout), workout type (HIIT, yoga, cardio, or strength), experience level (1 for beginner, 2 for intermediate, 3 for advanced), and calories burned.[5]

Before model training, the dataset was examined for data quality. There were no missing values or duplicate entries. Data types were analyzed, revealing that most features were numeric (int64 or float64), while Gender and Workout type were categorical (object) and required conversion. These categorical variables were label encoded, where Gender was mapped as Male = 0 and Female = 1, and Workout type as Yoga = 0, HIIT = 1, Cardio = 2, and Strength = 3. Additionally, two new features were engineered to capture workout intensity.

$$\text{Maximum BPM Range} = \text{Maximum BPM} - \text{Resting BPM} \quad (1)$$

$$\text{Average BPM Range} = \text{Average BPM} - \text{Resting BPM} \quad (2)$$

These new features provided better insight into how heart rate fluctuations during exercise relate to energy expenditure.

A correlation matrix was created to explore relationships between all variables. This analysis helped confirm which features were most strongly related to calories burned and supported the decision to exclude some redundant features. The results are presented in Figure 1.

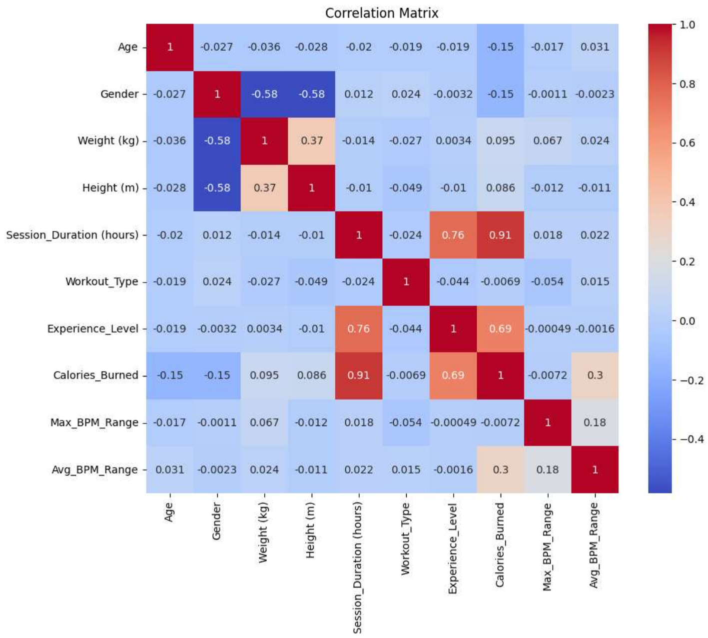

Figure 1: Correlation Matrix

As a result, the feature matrix X included all features except Calories Burned (target variable), Maximum BPM, Average BPM, and Resting BPM, since their effects were already captured by the BPM range features. The target variable y was defined as Calories Burned. The dataset was then split into training and testing sets, with 80% of the data used for training and 20% for testing, to evaluate the models on unseen data.

Three regression models were used to train and test the dataset: Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. For each model, performance was evaluated using the R^2 score, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Cross-validation was also applied to assess the consistency of the models. Additionally, feature coefficients were analyzed in the linear model, feature importance was reviewed in the Random Forest model, and the Gradient Boosting model yielded the highest prediction accuracy among the three. All the performance results from these models were compiled and presented in Table 1, which shows a comparative analysis of the models based on R^2 score, RMSE, MAE, and cross-validation scores, allowing for easier evaluation and selection of the best-performing model.

Table 1: Performance of Models

Model	**Metrics**			
	R^2 score	***RMSE***	***MAE***	***Mean Cross Validation Scores***
Linear Regression	0.9667	52.71	41.55	0.955
Random Forest Regressor	0.9547	61.50	47.74	0.946
Gradient Boosting Regressor	0.9779	42.92	34.13	0.9711

TUNE HYPERPARAMETERS AND VALIDATION OF THE MODEL

To optimize model performance, hyperparameter tuning was performed on both the Random Forest Regressor and Gradient Boosting Regressor using GridSearchCV with 5-fold cross-validation. The tuning process involved systematically searching through combinations of key hyperparameters specific to each model. For Random Forest, parameters such as the number of estimators, maximum depth, and minimum samples required for splits and leaves were explored. For Gradient Boosting, the tuning included variations in the number of estimators, learning rate, maximum depth, and subsample ratio. After tuning, both models were evaluated on the test dataset, and their performance was measured using R^2 score and RMSE. The improvement compared to the untuned models was also calculated. The cross-validation score provided an estimate of the model's generalization capability. A detailed summary of the tuned performance metrics and improvements for each model is presented in Table 2.

Table 2: Performance of Tuned models

Model	**Metrics**			
	R^2 score	***RMSE***	***Improvement***	***Cross Validation Score***
Tuned Random Forest Regressor	0.9547	61.50	0.0000	0.9464
Tuned Gradient Boosting Regressor	0.9790	41.87	0.0011	0.9720

RESULTS AND DISCUSSION

Linear Regression, while providing a reasonable baseline, is constrained by its assumption of linear relationships and additive feature effects. It exhibited relatively low generalization performance compared to the ensemble models. In contrast, the Random Forest Regressor captured non-linear dependencies more effectively through ensemble averaging of decision trees. Gradient Boosting Regressor demonstrated the highest predictive capacity, leveraging sequential learning and gradient-based optimization to minimize residual error at each iteration.

Random Forest and Gradient Boosting Regressors were both tuned using GridSearchCV with 5-fold cross-validation to identify optimal hyperparameter combinations. This included adjusting the number of estimators, tree depth, and other structural parameters relevant to each algorithm. The tuning process led to improved performance in the Gradient Boosting model, while no significant improvement was observed in the other model. The tuned Gradient Boosting Regressor achieved the best overall results and demonstrated strong consistency across cross-validation folds, confirming its robustness and ability to generalize. A comparative summary of tuned and untuned model performance, including cross-validation statistics and observed improvements, is provided in Table 2.

Figure 2 presents a scatter plot of actual versus predicted calorie values for the best performing model, the tuned Gradient Boosting Regressor. The red dashed diagonal represents the ideal reference where predicted values equal actual outcomes. The high degree of alignment between the data points and the reference line reflects the model's ability to generalize well across the test dataset, with minimal bias and variance.

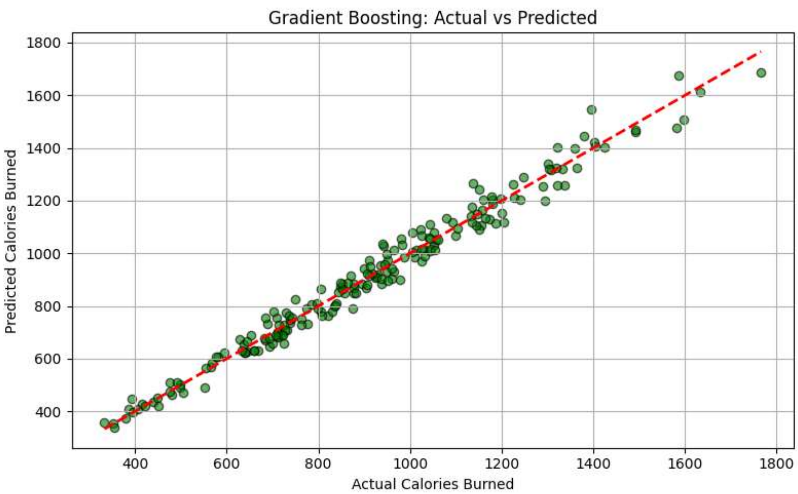

Figure 2: Actual vs predicted values using tuned Gradient Boosting Regression model

An example input was passed through the tuned Gradient Boosting model, which successfully returned a calorie prediction, demonstrating the model's suitability for practical use in fitness tracking.

CONCLUSION

This study successfully applied supervised machine learning techniques to predict calories burned during gym sessions using biometric and workout-related features. Among the models evaluated, Gradient Boosting Regressor demonstrated the best performance in terms of accuracy and generalization. Hyperparameter tuning further enhanced model effectiveness, confirming the importance of model optimization. The results indicate that ensemble methods, particularly Gradient Boosting, are well-suited for complex regression tasks involving non-linear relationships. With its high predictive accuracy and reliability, the final model has strong potential for integration into fitness applications, enabling personalized tracking and data-driven health insights.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Eng. S.U. Dampage, our lecturer, for the valuable guidance and continuous support provided throughout this project. His insightful explanations of real-world applications and consistent encouragement greatly motivated us to explore practical aspects of machine learning and apply them effectively.

REFERENCES

[1] J. S. A. Tan, Z. Che Embi, and N. Hashim, "Comparison of Machine Learning Methods for Calories Burn Prediction," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 182–191, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.12.

[2] G. Acoba, J. N. Mindoro, M. A. F. Malbog, J. A. B. Susa, and J. S. Gulmatico, "Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm," in *ICPC2T 2022 - 2nd International Conference on Power, Control and Computing Technologies, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICPC2T53885.2022.9776710.

[3] R. Miller, "How many calories does yoga burn?", *Healthline*, Aug. 24, 2020. [Online]. Available: https://www.healthline.com/health/how-many-calories-does-yoga-burn

[4] V. Alakhorasani, "Gym Members Exercise Dataset," *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset

[5] S. Supulsara, "Calories Burned Prediction (Project Code)," *GitHub*, Jul. 2025. [Online]. Available: https://github.com/sandali6723/calories-burned-prediction.git