Current Integrative Engineering
Volume 1, Issue 1, 29-36
DOI: 10.59762/cie570390541120231031130323

Calories Burnt Prediction Using Machine Learning Approach

Mohammad Tarek Aziz*, Taohidur Rahman, Renzon Daniel Cosme Pecho, Nayeem Uddin Ahmed Khan, Akba Ull Hasna Era, MD. Abir Chowdhury
Rangamati Science and Technology University, Bangladesh.

ABSTRACT
Calorie burnt prediction by machine learning algorithm" aim to predict the number of calories burnt by an individual during physical activity using machine learning techniques. We collected a dataset that includes features such as heart rate, body temperature, and duration of activity. We used various machine learning models, including XGBoost, linear regression, SVM and random forest, to predict calorie burn based on 15,000 records with seven features. The results indicate that the XGBboost model can accurately predict calorie burn with a minimum mean absolute error of calories. This work contributes to the growing body of research on using machine learning for health and fitness applications and has potential implications for personalized health coaching and wellness tracking.
The highest accuracy of training and testing is gained by the XGBboost model with 99.67% with mean absolute error is almost 1.48%.

Keywords: Machine learning algorithm, calories, health and fitness applications, XGBboost model.

Introduction
Background
Calorie is a unit of hear energy. Health and fitness are becoming increasingly important to individuals and society as a whole. As people seek to live healthier lifestyles, they are turning to wearable devices and fitness trackers to monitor their physical activity and track their progress. One important metric that these devices track is the number of calories burnt during physical activity. Accurately predicting calorie burn can help individuals set and achieve fitness goals and can also inform health coaching and wellness tracking programs [1]. The motivation for this research is to develop a machine-learning model that can accurately predict calorie burn during physical activity. This has potential applications in a range of settings, including personalized health coaching, fitness tracking, and wellness programs. By developing an accurate calorie burn prediction model, we can help individuals make more informed decisions about their physical activity and improve their overall health and well-being [2]. Although there has been some research on predicting calorie burn using machine learning techniques, there is still a significant gap in the literature. Most existing studies have focused on predicting calorie burn for specific types of physical activity or in specific populations. There is a need for more generalizable models that can accurately predict calorie burn across a range of physical activities and individuals [3]. The main objectives of this studies are: To collect data on physical activity and calorie burn from a variety of sources, including fitness trackers and wearable devices. Need to preprocess and clean the data to ensure accuracy and consistency. To develop a range of machine learning models to predict calorie burn, including linear regression,

Mohammad Tarek Aziz
Rangamati Science and Technology University
Bangladesh
tarekaziz4288@gmail.com

random forest, and support vector machines. To compare the performance of these models and identify the most accurate model for predicting calorie burn. To interpret the results and draw conclusions about the effectiveness of using machine learning for predicting calorie burn [4]. The scope of this research is to develop a machine learning model for predicting calorie burn during physical activity. We will collect data from a variety of sources and preprocess it to ensure accuracy and consistency. We will use several different machine learning models to predict calorie burn and compare their performance. The study will focus on a range of physical activities, including walking, running, and cycling. The study will not examine the impact of other factors such as diet or sleep on calorie burn [5]. In the discussion, section 2 provides a comprehensive review of the existing literature on machine learning-based calorie burnt prediction. In section 3, we detail the methodology adopted for data collection, preprocessing, feature engineering, and model development. Section 4 outlines the results of our experiments and compares the performance of various machine learning models. In section 5, we discuss the implications of our findings and identify potential areas for future research. Finally, in section 6, we summarize the key conclusions of our study and provide recommendations for both researchers and practitioners.

Literature Review:

Machine learning algorithms have gained widespread use in recent years to predict calorie burn during physical activity. These studies often collect physical activity data and other relevant variables such as heart rate, age, and gender from fitness trackers, mobile applications, and wearable devices. This section provides an overview of some of the critical studies in this area.

Sathiya T et al. [4] discussed to predict user's calorie and applied CNN model to classify food items from the input image. They also used image processing techniques such as deep learning model and their model provide 91.65% accuracy in predicting user's calorie from input image.

Sona P Vinoy illustrates to predict calorie burn during the workout et al. [6] used machine learning algorithms such as XGBboost regressor and Linear regression models to find out calorie burnt in physical activities. Their mean absolute error value is almost 2.71 in XGB regressor and 8.31 for linear regression. They used 7 attributes such as age, height, weight, duration, heart_rate, body_temp and calorie. Their dataset was in 15000 CSV with 7 attributes. They did not mention their model accuracy.

Suvarna Shreyas Ratnakar et al. [7] discussed how to predict calories burnt from physical activities. They used the XGB boost Machine learning algorithm to predict it including 15,000 raw dataset and their mean absolute error value is 2.7 and model accuracy is not mentioned. Rachit Kumar Singh et al. [8] illustrated their method to predict calorie burn using machine learning techniques. In their work, logistic regression, linear regression and lasso regression models were used but they didn't mention about mean error absolute value, dataset and model accuracy.

Marte Nipas et al. [9] discussed how to predict burned calories using a supervised learning algorithm. They used a Random forest algorithm and gained 95.77% model accuracy. They also used the iterative method to find out the appropriate output from an input. Their work is almost better than other recent work.

Gunasheela B L et al. [10] discussed their techniques to predict calorie from input images. They used some digital image processing techniques such as image acquisition, RGB conversion, feature extraction and image enhancement so on. They segmented input images and used techniques and then combined segmented images, finally calorie predicted.

KR Westerterp et al. [11] discussed how to determine energy expenditure by body size and body compositions and food intake and physical activity. He used body size and body compositions and some statistical techniques to evaluate calorie expenditure.

In summary, these studies demonstrate the potential for machine learning algorithms to predict energy expenditure accurately during physical activity. However, there is still a need for models that can accurately predict energy expenditure across various physical activities and individuals.

In the next section, we describe the methodology used in this study to address this research gap.

Methodology:

This study aims to predict the calorie burn during physical activity using machine learning models. The basic working flow diagram is illustrated in Fig. 1

3.1 Dataset Description

Data collection is an essential process in any machine learning project, as the quality of the data used has a significant

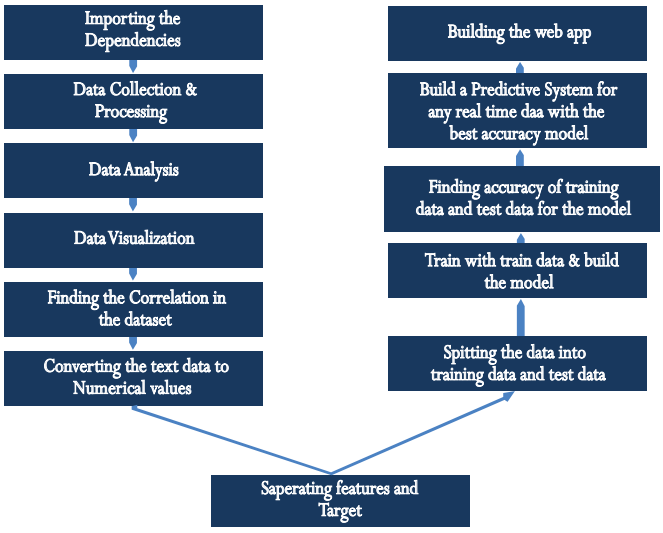

figure no 1

impact on the performance of the resulting model. In this research, the dataset was collected from Kaggle [12], a popular platform for data scientists and machine learning practitioners to access and share datasets. Once the dataset was collected, it was uploaded to Google Colab, a cloud-based platform for data analysis and machine learning. Google Colab. In this work, the dataset contained over 15,000 records and 7 variables.

3.2 Dataset Preprocessing

We preprocessed the data by removing missing values and outliers. Because, preprocess datasets are appropriate for applying into the algorithm for training and testing. We split the data into a training set (80% of the data) and a test set (20% of the data) for model training and evaluation [13].

3.3 Evaluation of the Performance of Machine Learning Models

We evaluated the performance of four machine learning models: support vector machine (SVM), random forest, linear regression, and XGBoost regression [14].

3.4 Comparison of Feature Selection with Individual Evaluators

We contrasted the performance of models that used all features to models that used feature selection methods such univariate feature selection and recursive feature elimination. To determine the relevance of the feature, we employed the correlation matrix [15].

3.5 Deriving Key Features

We derived key features by analyzing the feature importance scores from the models. The key features included heart rate, duration, and temperature. In summary, this study used data to predict calorie burn during physical activity using machine learning models. We preprocessed the data, evaluated the performance of the models, and also build a predictive system for any real time data. The results of this study are described in the next section.

3.6 Data Visualization

Dataset is visualized at Fig. 2; where we found two category datasets. These are: male and female in X axis and dataset counted in Y axis. In fig. 3, height vs density is illustrated; where the highest density is taken 0.025 to Y axis and height is last in 220 at the X axis. We can follow data visualization from image processing also [14].

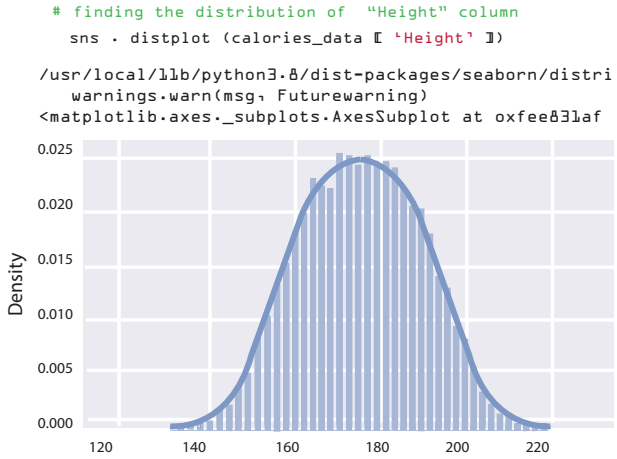

Figure. 2: Plotting the gender column in the count plot (data visualization)

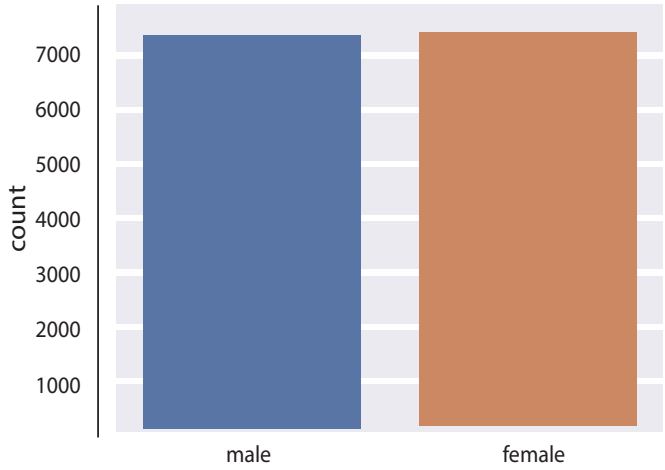

Figure. 3 finding the distribution of Height column

3.7 Correlation in the Datasets:

Correlation in the datasets among features is illustrated in fig. 4; it is indicated that interrelation among features of used data [15]

3.8 Building the Web App:

Actually, after building the web app; it predict the amount of calorie burnt based on input

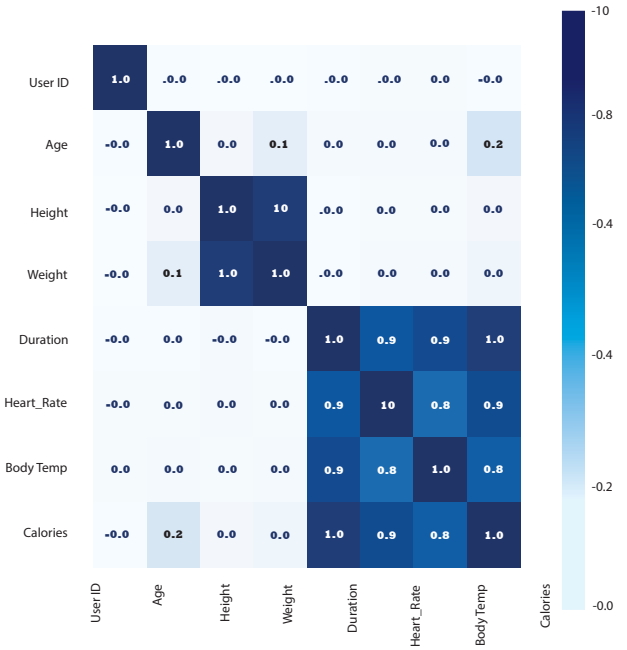

Figure 4 Construction of a heatmap to understand the correlation of the features

features. If we give 7 inputs, then the app can predict calorie burnt amount. The features are: gender, age, height, weight, duration,

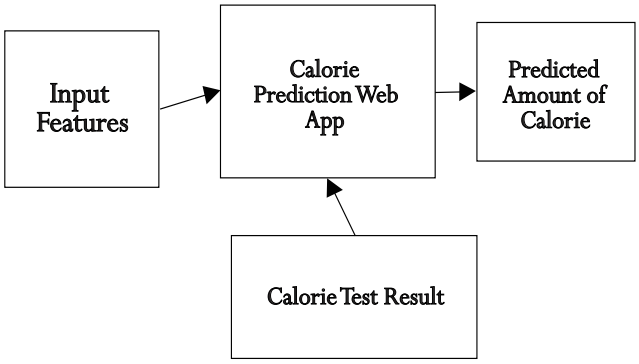

Figure 5 : How to predict calorie burnt by web App

heartrate and body temperature. After giving these features, our app will produce burnt calorie amount automatically. It can work like fig. 5

Results and Discussion:

In the result part, we will discuss about models accuracy for training and testing, different types of errors, bar chart of accuracy for different algorithm, bar chart of evaluation metrics of different algorithm, web app predicted results and finally comparison our work to recently done work.

The training and testing accuracy over same dataset for different model is shown in Table 4.1; where we can see that the highest accuracy is gained by XGBoost algorithm and lowest accuracy is gained by SVM. That's why, we chosen XGBoost model to build up web app for predicting calorie burnt amount. This app can find out the calorie burnt amount based on XGBoost algorithm at back end. So, by using this app we can get calorie burnt amount from physical activities with seven features [16].

Models	Training Accuracy	Testing Accuracy
SVM	19.71%	12.50%
Random Forest	100%	14.27%
Linear Regression	70.78%	72.21%
XGBoost	99.67%	99.63%

Table 4.1 Training and testing accuracy of different algorithm over same dataset

20%. Lowest accuracy is for SVM, less than 20% for both training and testing. On the other hand, we got highest polygon for XGBoost. Means that, it is appropriate

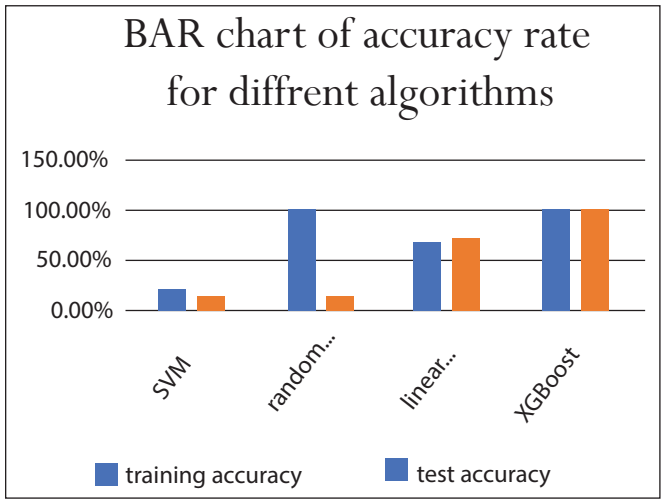

Figure 6 BAR chart of accuracy rate for different algorithm

and testing. It is satisfactory for our system. Where we see that training accuracy is highest for random forest but test accuracy is not satisfactory.

From the Table 4.2 illustrated different types of errors in Linear Regression and XGBoost models. Here lowest mean squared error is found in XGBoost regression model and it is appropriate for our calorie prediction system.

Fig. 7 represents that BAR chart of evaluation matrices of different algorithms. Where Linear regression and XGBoost models polygon are shown. Highest frequency polygon found for linear regression. Means that it has much error than from XGBoost regression. Every model errors has four parts. These are: Mean absolute error, Root mean absolute error, R-squared error, and Mean square error. In contrast, linear regression has significant inaccuracy. That is why we decided to design the app using XGBoost. These four typical indicators are frequently employed to assess the effectiveness of a regression model.

Model	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R-squared Score
Linear Regression	130.087	11.405	8.385	0.966
XGBoost Regression	4.534	2.129	1.480	0.998

Table 4.2 Score of different types of errors in model

1. Mean Squared Error (MSE): Between the expected and actual numbers, this calculates the average squared difference. The projected and actual values diverge more, as shown by a higher MSE.

2. Root Mean Squared Error (RMSE): A more understandable number in the same units as the target variable is provided by this, which is the MSE's square root.

3. Mean Absolute Error (MAE): This calculates the typical absolute difference between the expected and observed values. MSE is more sensitive to outliers than MAE.

4. R-squared (R2) score: This gauges how much of the target variable's variance the model is capable of explaining. A perfect fit is indicated by a score of 1, which goes from 0 to 1.

The MSE in the instance of linear regression is 130.09, suggesting that the average squared difference between the predicted and actual values is fairly significant. The average difference between the predicted and actual values is approximately 11.41 units, according to the

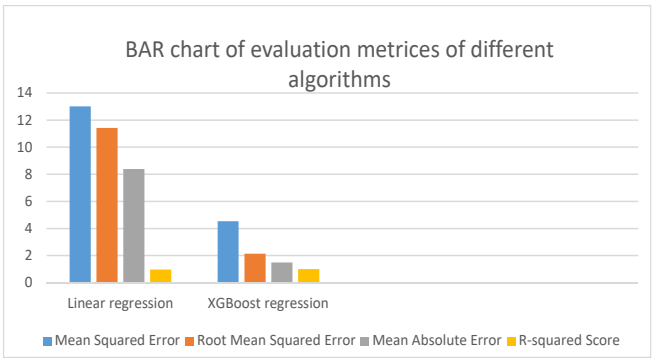

Fig.7 BAR chart of evaluation matrices of diffrent algorithms

RMSE, which is 11.41. The MAE is 8.39, which means that there is an 8.39 unit average absolute difference between the expected and actual values. Finally, the R2 score of 0.97 shows that the model explains 97% of the variance in the target variable, which is regarded as a good match.

Conversely, in the instance of XGBoost, the MSE is 4.53, suggesting that the average squared difference between the predicted and actual values is fairly small and the model has a strong fit. The RMSE is 2.13, which means that the average deviation between the predicted and actual values is roughly 2.13 units, which is a rather accurate result.

actual values. Lastly, the target variable's variance is explained by the model to a great extent (R2 = 0.9988), accounting for 99.88% of the variance in the target variable. The XGBoost regression model appears to have performed admirably on the provided dataset, according to these metrics.

1. Mean Squared Error (MSE): Between the expected and actual numbers, this calculates the average squared difference. The projected and actual values diverge more, as shown by a higher MSE.
2. Root Mean Squared Error (RMSE): A more understandable number in the same units as the target variable is provided by this, which is the MSE's square root.
3. Mean Absolute Error (MAE): This calculates the typical absolute difference between the expected and observed values. MSE is more sensitive to outliers than MAE.
4. R-squared (R2) score: This gauges how much of the target variable's variance the model is capable of explaining. A perfect fit is indicated by a score of 1, which goes from 0 to 1.

The MSE in the instance of linear regression is 130.09, suggesting that the average squared difference between the predicted and actual values is fairly significant. The average difference between the predicted and actual values is approximately 11.41 units, according to the RMSE, which is 11.41. The MAE is 8.39, which means that there is an 8.39 unit average absolute difference between the expected and actual values. Finally, the R2 score of 0.97 shows that the model explains 97% of the variance in the target variable, which is regarded as a good match.

Conversely, in the instance of XGBoost, the MSE is 4.53, suggesting that the average squared difference between the predicted and actual values is fairly small and the model has a strong fit. The RMSE is 2.13, which means that the average deviation between the predicted and actual values is roughly 2.13 units, which is a rather accurate result. The MAE is 1.48, indicating an average absolute difference of about 1.48 units between the expected and actual values. Lastly, the target variable's variance is explained by the model to a great extent (R2 = 0.9988), accounting for 99.88% of the variance in the target variable. The XGBoost regression model appears to have performed admirably on the provided dataset, according to these metrics.

Fig. 8 shows that calorie prediction web-based app. It is an real life app to find out burnt calorie amount by giving some input features such as gender, age, height, weight, duration, heart rate and body temperature. After giving this input, this app will predict calorie burnt amount. In the discussion, we can compare our work to other researcher's existing work. In the comparison, we can see that our work is superior to other. For analyzing, see the comparison table below.

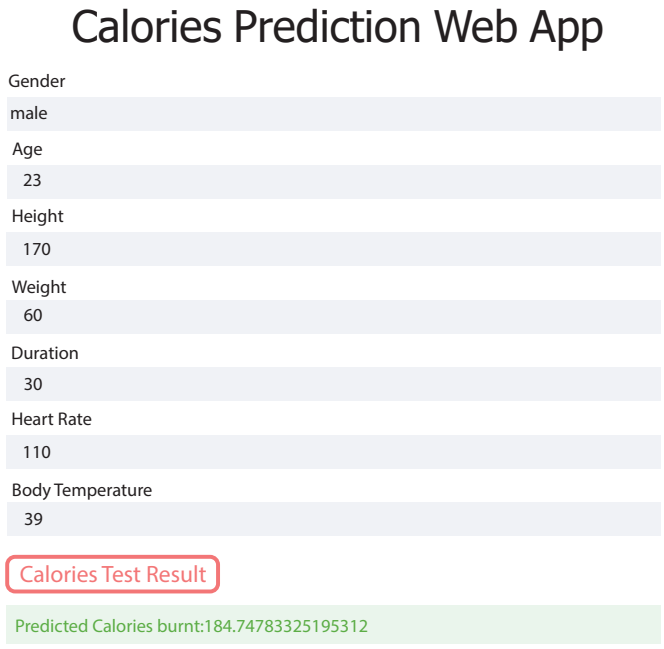

Figure. 8 Calories Prediction Web App

Conclusion and Future Plan:

The main objective of this study was to create a precise machine learning model that could predict a specific outcome variable based on a series of characteristics. This was achieved through the collection and preparation of a dataset, as well as testing the effectiveness of various machine learning models and feature selection techniques. The findings indicated that the XGBoost model demonstrated superior performance compared to the other models in terms of accuracy and other relevant metrics. This suggests that the XGBoost model could be a valuable tool for

predicting similar outcome variables based on similar datasets. The key features derived from the feature selection and evaluation process were identified and discussed in terms of their importance and relevance in predicting the target variable. These key features were also related to the problem domain, providing insights and implications for potential applications. However, the study also had some limitations, such as the limited size of the dataset and the possibility of overfitting. Future research could address these limitations and further improve the performance of the models and feature selection approaches. Overall, this study contributes to the field of machine learning and provides practical implications for the problem domain.

Article	Title	Datasets	Method	Accuracy	Mean absolute error
[4]	PREDICTION OF USER'S CALORIE ROUTINE USING CONVOLUTIONAL NEURAL NETWORK	Random Images from Google	CNN	91.65%	-
[6]	Calorie Burn Prediction Analysis Using XGBoost Regression and Linear Regression Algorithms	15,000 records with 7 variables	XGBoost, Linear Regression	-	2.71(XGBoost), 8.31(Linear Regression)
[7]	Calorie Burn Prediction using Machine Learning	15,000 records with 7 variables	XGBoost	-	2.7
[8]	Calories Burnt Prediction Using Machine Learning	-	Logistic Regression, Linear Regression and Lasso Regression	-	-
[9]	Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm	-	Random Forest	95.77%	-
[10]	CALORIES PREDICTION BASED ON FOOD IMAGES	-	Digital Image Processing Techniques	-	-
Proposed Work	Calories Burnt Prediction Using Machine Learning Approach	15,000 records with 7 variables	XGBoost, SVM and Linear Regression	99.67%	1.48%

Table 4.3 Comparative analysis proposed work to existing researchers

References

[1] D. Bubnis, "Calculating how many calories are burned in a day," Medical News Today, 01 January 2020. [Online]. Available: https://www.medicalnewstoday.com/articles/319731.
[2] K. S. University, "Burning more calories is easier when working out with someone you perceive as better," 26 November 2012. [Online]. Available: https://www.sciencedaily.com/releases/2012/11/121126130938.htm.
[3] B. K. Tingley, "The New Science on How We Burn Calories," The New Work Times Mag

azizne, 14 September 2021. [Online]. Available: https://www.nytimes.com/2021/09/14/magazine/calories-weight-age.html.
[4] S. T and V. K, "PREDICTION OF USER'S CALORIE ROUTINE USING CONVOLUTIONAL NEURAL NETWORK," International Journal of Engineering Applied Sciences and Technology, vol. 5, no. 3, pp. 189-195, 2020.
[5] G. vijayalakshmi and T. Sridurga, "COMPARING MACHINE LEARNING ALGORITHMS FOR PREDICTING CALORIES BURNED," JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR), vol. 10, no. 3, pp. 519-527, March 2023.
[6] S. P. Vinoy and B. Joseph, "Calorie Burn Prediction Analysis Using XGBoost Regressor and Linear Regression Algorithms," in Proceedings of the National Conference on Emerging Computer Applications (NCECA), Kottayam, 2022.
[7] S. S. Ratnakar and V. S, "Calorie Burn Predection using Machine Learning," International Advanced Research Journal in Science, Engineering and Technology, vol. 9, no. 6, pp. 781-787, June 2022.
[8] R. K. Singh and V. Gupta, "Calories Burnt Prediction Using Machine Learning," International Journal of Advanced Research in Computer and Communication Engineering, vol. 11, no. 5, May 2022.
[9] M. Nipas, A. 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 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, March 2022.
[10] R. S. Biyani and M. S. Nandini, "CALORIES PREDICTION BASED ON FOOD IMAGES," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 8, pp. 2122-2125, August 2020.
[11] K. Westerterp, "Control of energy expenditure in humans," European Journal of Clinical Nutrition, vol. 71, pp. 340-344, 30 November 2016.
[12] "calories_burnt _data," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/sandeepgauti/calories-burnt-data.
[13] K. K. Al-jabery, T. Obafemi-Ajay, G. R, Olbricht and D. C. W. II, "Computational Learning Approaches to Data Analytics in Biomedical Applications," ACADEMIC PRESS, 2020, pp. 7-27.
[14] K. Nighania, "Various ways to evaluate a machine learning model's performance," Towards Data Science, 30 December 2018. [Online]. Available: https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15.
[15] T. Z. Phyu and N. N. Oo, "Performance Comparison of Feature Selection Methods," in MATEC Web of Conferences, EDP Sciences, 2016.
[16] M. T. AZIZ, J. SIKDER, T. RAHMAN, A. D. D. MUNDO, S. F. FAISAL and N. U. A. KHAN, " COVID-19 DETECTION FROM CHEST X-RAY IMAGES USING DEEP LEARNING," The Seybold Report, vol. 17, no. 11, p. 706–718, 2022.
[17] J. N. P. Ling, "Heatmap For Correlation Matrix & Confusion Matrix | Extra Tips On Machine Learning," MLearning.ai, 19 February 2022. [Online]. Available: https://medium.com/mlearning-ai/heatmap-for-correlation-matrix-confusion-matrix-extra-tips-on-machine-learning-b0377cee31c2.
[18] X. Wang, D. Fu, Y. Wang, Y. Guo and Y. Ding, "The XGBoost and the SVM-based prediction models for bioretention cell decontamination effect," Arabian Journal of Geosciences , vol. 669, no. 14, 6 April 2021.

Recieved: May-4-2023 Revised: Aug-13-2023 Accepted: October-22-2023