International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.51
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Application of Ensemble Learning and Feature Selection Models in Energy Expenditure Estimation of Fitness Tracking Devices

Hefan Wei*

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

ABSTRACT

The use of fitness tracking devices to monitor energy expenditure (EE) is becoming increasingly popular among consumers. However, the accuracy of EE estimation still needs to be improved due to the lack of accurate data filtering and more efficient data analysis models. To address this issue, we designed an energy expenditure estimation (EEE) model based on smart tracking devices. This paper proposes an algorithm for data reconstruction called PC-DF-RFECV: firstly, Pearson correlation analysis (PC) is used for initial data screening, then Feature Derivation (DF) is applied to reconstruct new features, and finally, the Recursive Feature Elimination algorithm based on cross-validation (RFECV) is employed to select the new features. Furthermore, the EEE model adopts a weighted average (Wei-Voting) strategy to enhance the robustness of the model by integrating the predictions of multiple base learners (XGBoost, LightGBM, Random Forest). During algorithm optimization, a Bayesian hyperparameter optimization technique is employed to fine-tune the model's hyperparameters. Empirical findings indicate that the EEE model, when applied to the dataset collected from motion tracking devices, attains an score of 0.9985, of 1.033, and of 0.865. These results demonstrate superior performance compared to current state-of-the-art research and conventional algorithms including random forests, gradient boosting, and bagging.

KEYWORDS

Energy expenditure; Machine learning; Feature selection; Wearable technology; Hyperparameter optimization

1. INTRODUCTION

Accurate measurement of EE is essential for individuals to understand their calorie consumption and improve their personal health. Energy consumption can be measured using methods such as calorimetry and bioresistance analysis experiments [1]. Nevertheless, traditional approaches prove to be both resource-intensive and time-consuming, while the intricate structure of the equipment poses challenges for sustained real-time measurements. The use of smart devices combined with data-driven methods can solve this problem. Wearable smart devices, such as Fitbit, Apple Watch, and Nike + FuelBand, can integrate these models to monitor various physiological outputs. Validity and reliability studies have demonstrated the usefulness of these devices [2] for monitoring human activity (HA) [3] and health conditions, including sleep quality [4] heart rate [5], and weight management [6]. However, it lacks algorithms to process physiological data in deeper research, resulting in deficiencies in predictive assessments of EE. Thus, this study proposes using machine learning algorithms to identify the relationship between EE and features, select optimal physiological features, and map input data to corresponding outputs for EE prediction. The primary focus of this study is to address this issue and improve the accuracy of EEE [7].

In the physiological data analysis phase, machine learning plays a crucial role in analyzing human body indicators using data acquired from wearable smart devices. Wang et al. [8] conducted a study on daily living aids for the elderly, where they combined environmental sensors and wearable sensors to recognize human activities. Guo et al. [9] developed a SmartBand device to collect data and employed the XGBoost algorithm for classifying the health level of adolescents. Sattar et al. [10] used smartphone sensors to sample low-frequency data and utilized artificial neural networks to model EEE. Yuan et al. [11] proposed the coronavirus mask protection algorithm, a protective biomimetic algorithm aimed at preventing the transmission of the novel coronavirus. This algorithm mathematically simulates human self-protection behavior by considering correct mask wearing and safe social distancing as parameters. The aforementioned studies demonstrate the efficacy of traditional machine learning methods in extracting valuable insights within the domain of human health.

Ensemble learning methods like Bagging and Boosting have captured considerable interest within the realm of data analysis. These approaches are also prevalent in medical applications for disease identification and predicting engineering problems' performance. The Boosting algorithm combines several weak machine learners by first training the individual learners with the training dataset. It then updates the sample weights based on the weak learner's learning error rate, increasing the weights of training samples where the weak learner performs more efficiently. This strategy aims to achieve optimal learning outcomes. Yang et al. [12] proposed a novel integration algorithm called Rotation-Flexible AdaBoost designed to enhance integration diversity by rotating feature axes and employing a random subspace method for each bootstrap sample. The authors applied this approach to 30 binary classification problems, achieving excellent classification results. Ren et al. [13] proposed an AdaBoost double-layer learner, which combined random forest as weak classifiers with genetic algorithm-assisted improvement, to predict the oil temperature of tunnel boring machines in order to avoid malfunctions caused by excessive oil temperature. These studies highlight the potential benefits of incorporating innovative techniques and multiple algorithms to improve model performance in various real-world applications.

According to the "No Free Lunch" theorem [14], a single algorithm cannot effectively solve all engineering problems. The ensemble regressors showed better performance than the single regressors, but the prediction results varied due to the different prediction errors of each underlying ensemble regressor. The weighted voting method outperforms the base learner and the relative majority voting method by configuring appropriate weights for the base learner. Kim et al. [15] highlight the superiority of weighted voting over the simple majority voting method. Furthermore, to enhance learning efficiency, the TPE algorithm is employed to optimize the hyperparameters of the base learner.

2. EXPERIMENTAL EQUIPMENT AND EXPERIMENTAL PARADIGM

Fitbit is a widely used commercial fitness aid that leverages a tri-axis accelerometer and altimeter to accurately recognize human movements, surpassing the limited accuracy of previous single-axis pedometers. The accelerometer quantifies the frequency and duration of human movement to estimate the user's energy expenditure (EE). The Fitbit activity tracker communicates with a mobile device through Bluetooth technology, facilitating the continuous collection and storage of data in a cloud-based database. The analysis process involves retrieving data from the cloud and applying machine learning algorithms. To facilitate this process, a conceptual framework for EE has been developed, which includes examples of digital phenotypes that can aid in identifying unique patterns of physical activity and energy expenditure. Figure. 1 illustrates the proposed model's flowchart, comprising four modules: data provisioning, feature processing, model execution, and model evaluation. Together, these modules form the EEE model.

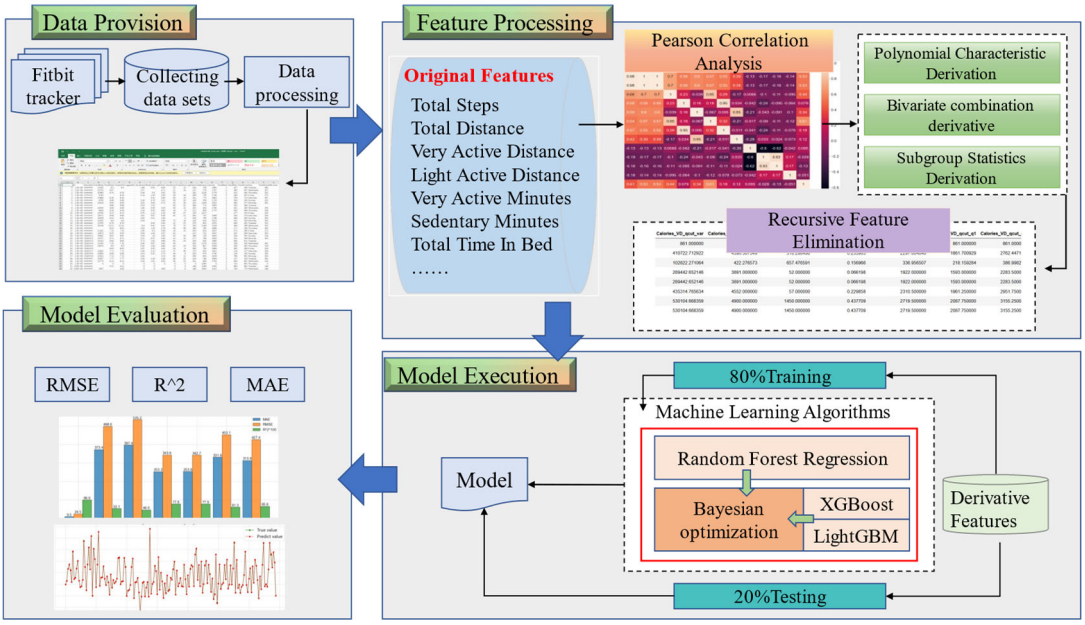

Figure 1. Overall flow and structure of the analysis process for estimating energy expenditure based on Fitbit devices

3. FEATURE ENGINEERING

3.1. Data Description and Pre-processing

The missing values in some fields were filled using a preferred strategy inferred from the business logic. The distances over which the volunteers were active at three different exercise intensities are shown in Figure 2. The distance volunteers moved at a given exercise intensity varied little from Monday to Friday. The activity distance on Saturday is higher than other days, and the activity distance on Sunday is much lower than other days. Correspondingly, the total activity distance of volunteers decreased significantly on Sunday and peaked on Saturday. This suggests that there is a pattern of movement in the volunteers. For such regular and locally missing data, a suitable interpolation function $f(x)$ can be established using the known points, and the function value $f(x_i)$ of the unknown point x_i can be found by the interpolation function $f(x)$. Therefore, it is more effective to choose to use the Lagrange Interpolation Polynomial method to fill in the missing data values with the obtained.

For n known points in the plane that are dissimilar, assume that the polynomial with degree $n-1$ is represented as $y = m_0 + m_1 x + m_2 x^2 + \ldots m_{n-1} x^{n-1}$, and bringing n points into the polynomial yields the following equation:

$$\begin{aligned} y_1 &= m_0 + m_1 x_1 + m_2 {x_1}^2 + \ldots m_{n-1} {x_1}^{n-1} \\ y_2 &= m_0 + m_1 x_2 + m_2 {x_2}^2 + \ldots m_{n-1} {x_2}^{n-1} \\ &\ldots \\ y_n &= m_0 + m_1 x_n + m_2 {x_n}^2 + \ldots m_{n-1} {x_n}^{n-1} \end{aligned} \tag{1}$$

The interpolation result $L(x)$ for missing values can be expressed as follows (Here the known points two days before and after the missing values are chosen to build the interpolation function):

$$L(x)=\sum_{i=1}^{n} y_i \prod_{j=1, j\neq i}^{n} \frac{x-x_j}{x_i-x_j} (0<n\leq 2) \tag{2}$$

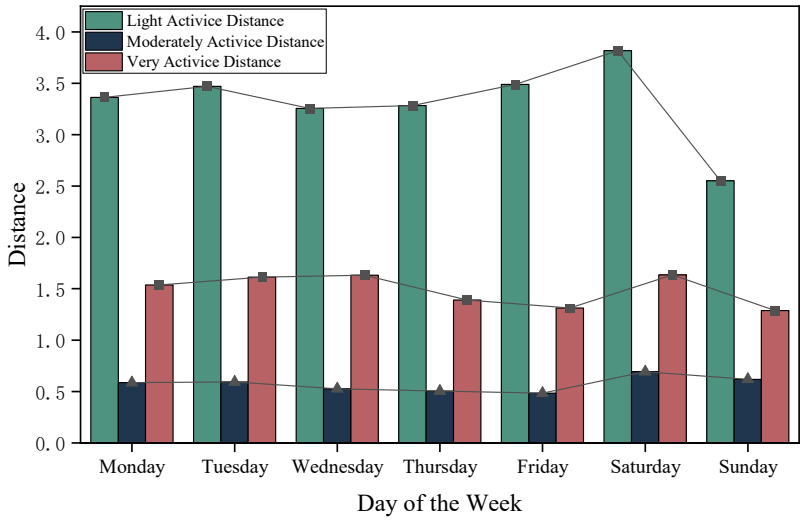

Figure 2. The daily exercise distance statistics corresponding to different intensity exercises.

3.2. Pearson Correlation strategy

The Pearson correlation coefficient quantifies the linear correlation between two variables, defined as equation 3. Where σ_X and σ_Y symbolize the standard deviation of the respective variables, $cov(X,\ Y)$ represents the covariance of the two variables, which is calculated using the equation 4.

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

$$\operatorname{cov}(X,Y)=\frac{1}{n-1}\sum_{i=1}^{i=n}(X_i-\bar{X})(Y_i-\bar{Y}) \tag{4}$$

Pearson correlation analysis is used to explain the potential impact of variable dimensions on the absolute value of covariance [16], assessing the linear association between two variables. The correlation coefficient ranges from -1 to 1. We calculated the Pearson correlation coefficient between the different variables in the original data A correlation coefficient of -1 signifies a complete negative linear correlation, 1 indicates a complete positive linear correlation, while 0 suggests no linear relationship between the variables. By analyzing the correlation between variable dimensions and the absolute value of covariance, we can understand how changes in variable dimensions affect the strength and direction of covariance. Additionally, this analysis helps filter out features in the original dataset that have weak correlations with the target variable. Calculations show that the duration of intense exercise (i.e., Very Active Distance) has a strong correlation with EE, while the distance-related features (i.e., Total Distance, Tracker Distance) also exhibit a notable association with EE.

3.3. Feature Derivation

Currently, the HA data collected through Fitbit often contains numerous raw features that lack meaningful interpretation, failing to reflect the crucial information related to EE in the dataset. The transformation or combination of certain features in the dataset may lead to high information value and significantly improve data quality. Feature derivation involves reconstructing existing big data features from the perspective of business and data to generate new features. In this paper, we propose two types of feature derivation approaches based on the characteristics of HA data and big data for HA data.

(1) In order to address the issue of high-dimensional sparse features in the HA dataset, which tend to lack meaningful information, statistical class indicators can be constructed to reflect their concentration trends and dispersion. Based on these statistics, new features can be derived, including mean, maximum, minimum, median, sample variance, sample sum, and quantile.

(2) Feature derivation based on feature meaning involves identifying features that have business relevance to each other and combining them to extract new features that align with the prediction scenario requirements. By leveraging business logic ideas, related features can be combined to create new, meaningful features.

3.4. RFECV

RFECV is a commonly used feature selection method that recursively removes the least significant features and employs cross-validation to assess model performance. This method aids in identifying the most relevant features for classification tasks and helps avoid overfitting. The use of cross-validation guarantees that the model's performance is unaffected by the selection of a specific subset of features. RFECV is implemented in two steps:

(1) RFE is a widely used feature selection technique in supervised learning. It iteratively constructs a learning estimator, assigns importance scores to features, removes the least important ones, and repeats until a desired number of features or performance is achieved. RFE efficiently identifies and removes irrelevant or redundant features, resulting in a more accurate and interpretable model. It has proven efficacy in various domains, making it a valuable tool in many applications.

(2) CV can assist in identifying the optimal number of features. For base model cross-validation, we first compute the average score when no features are removed, then compute the score for all combinations of n features that are removed, and repeat until we find the minimum number of features to remove, thus determining the optimal feature subset.

4. MACHINE LEARNING MODELS

4.1. Random Forest

Random forest, an ensemble algorithm employing the bagging technique, has shown comparable performance to boosting. This algorithm randomly selects samples with replacement from the training data and builds a decision tree model for each iteration. Subsequently, the predictions of multiple decision trees are combined using the majority voting principle to assign the prediction category for each sample. These predicted outcomes can be formulated as follows:

$$T(x) = \arg\max_{y} \sum_{n=1}^{N} g_n(x)(g_n(x) = y) \tag{5}$$

where $T(x)$ denotes the prediction outcome of the random forest, while x represents the feature vector $[x_1, x_2, x_3, ..., x_n]^T$. N signifies the count of decision trees in the random forest, and $g_n(x)$ stands for the prediction outcome of the nth decision tree.

4.2. XGBoost

Chen [17] developed XGBoost, a gradient boosting-based decision tree integration algorithm. This is an improved and refined version of the GBDT algorithm, which utilizes a second-order Taylor expansion of the loss function for optimization. To overcome overfitting, XGBoost incorporates a regularization term in the objective function to reduce variance. These enhancements have resulted in XGBoost's superior performance over GBDT. Figure. 3 shows the internal structure of the XGBoost algorithm, the initial predicted value is $F_0 = 0$, the model trains a regression tree with

residuals $y-F_n$ as labels, f_n as the residual prediction value, update the prediction of the model to $f_n=F_{n-1}+f_n$, f_n continues as the label of the next tree, and so on, until F_k is outputted to predict the result .

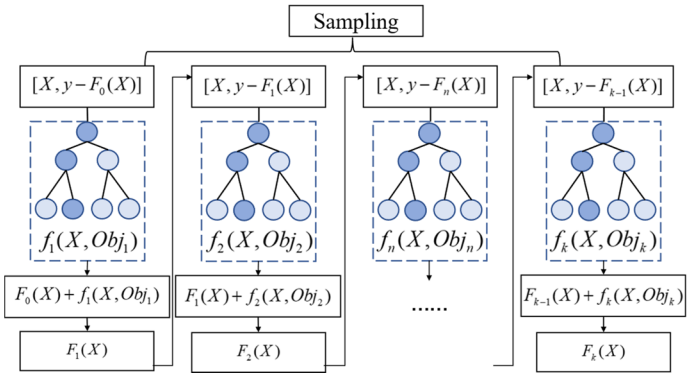

Figure. 3. Schematic diagram of the XGBoost and LightGBM algorithms.

XGBoost's objective function comprises two main components: a loss function L and a regularization term Ω, which regulates the complexity of the model.

$$Obj^{(k)}=\sum_{m=1}^{t}L\left(y_m,\hat{y}_m^{(k)}\right)+\sum_{m=1}^{k}\Omega\left(f_k\right) \tag{6}$$

The regularization term in the objective function includes the summed complexities of all k trees to prevent model overfitting. Each leaf node's weights also play a role in the regularization. The expression for the regularization term $\Omega\left(f_k\right)$ is provided by the following formula:

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{n=1}^{T}\omega_n^2 \tag{7}$$

where γ is the penalty term coefficient, T denotes the number of leaf nodes, and $\frac{1}{2}\lambda\sum_{n=1}^{T}\omega_n^2$ is the $L2$ parametrization.

After k iterations, the model is updated as the result of the previous $t-1$ iterations plus a new decision tree, so that the loss function is updated as:

$$Obj^{(k)}=\sum_{m=1}^{t}L\left(y_m,\hat{y}_m^{(k-1)}+f_k(x_m)\right)+\Omega\left(f_k\right) \tag{8}$$

Unlike GBDT, which relies solely on first-order derivatives for optimization, XGBoost employs the second-order Taylor expansion technique to optimize the cost function. The objective function's Taylor expansion is formulated as follows:

$$Obj^{(k)}\approx\sum_{m=1}^{t}[L(y_m,\hat{y}^{(k-1)})+g_m f_k(x_m)+\frac{1}{2}h_m f_k^2(x_m)]+\Omega(f_k) \tag{9}$$

The objective function's Taylor expansion in XGBoost incorporates both the first-order derivative g_m and the second-order derivative h_m of the loss function.

4.3. LightGBM

Ke et al. [18] introduced LightGBM, a gradient-boosting decision tree framework, which efficiently implements the GBDT algorithm using Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS). LightGBM employs a histogram-based decision tree approach to convert

continuous feature values into K integers and constructs a histogram of width K. During data traversal, the histogram gathers statistics of the discretized values to identify the optimal leaf partition point. Unlike XGBoost, LightGBM follows a leaf-wise growth strategy with a depth limit and selects only the node with the highest gain for splitting at each iteration, as illustrated in Figure 4. This approach reduces the search and splitting of nodes with lower gain, reducing overhead and enhancing efficiency. Overall, these optimizations lead to improved performance compared to other gradient-boosting decision tree frameworks.

Figure 4. The growth strategy of LightGBM.

Similar to the XGBoost algorithm, LightGBM aims at minimizing the following regularization objectives.

$$Obj^{(k)} = \sum_{m} L\left(y_m, \hat{y}_m^{(k)}\right) + \sum_{k} \Omega\left(f_k\right) \tag{10}$$

After k iterations, the objective function of the regression tree can be expressed as:

$$Obj^{(k)} = \sum_{m=1}^{t} L\left(y_m, f_{k-1}(x_m) + f_k(x_m)\right) + \sum_{k} \Omega\left(f_k\right) \tag{11}$$

Further, Newton's method is used to LightGBM to fast approach the objective function, which can be expressed as:

$$Obj^{(k)} \cong \sum_{m=1}^{i} [g_m f_k(x_m) + \frac{1}{2} h_m f_k^2(x_m)] + \sum_{j} \Omega(f_j) \tag{12}$$

where the first- and second-order derivatives are represented by g_m and h_m can be expressed as:

$$g_m = \partial_{F_{k-1}(x_m)} \varphi(y_m, F_{k-1}(x_m)) \tag{13}$$

$$h_m = \partial^2_{F_{k-1}(x_m)} \varphi(y_m, F_{k-1}(x_m)) \tag{14}$$

5. PROPOSED ALGORITHM

5.1. Feature Selection (PC-DF-RFECV)

A high-level feature selection scheme, named PC-DF-RFECV, has been proposed to extract key features related to EE from data samples. Initially, Pearson correlation analysis is applied to extract features with a high correlation to the target value level from the original features. However, this only eliminates the features that are not correlated in terms of numerical relationships. Therefore, we introduced the DF-RFECV method to reprocess the features: Among them, DF (Derivative Features) derives a large number of features, and RFECV does screening on a large amount of derived data to remove redundancy between features. Experimentally, this method proved to be very effective in feature selection.

5.1.1. Bivariate Polynomial Feature Derivation

Prior to deriving binary polynomial features, it was necessary to bin different intervals of eigenvalues. In the screening process of the features, those with Pearson correlation coefficients greater than 0.4 were classified as strongly correlated features, while the remaining features were categorized as weakly correlated features. The results of the classification are shown in Table 1.

Table 1. Partitioning of strongly/weakly correlated subsets.

Strongly related subsets	Weak related subsets
Total Steps	Moderately Active Distance
Total Distance	Light Active Distance
Very Active Distance	Fairly Active Minutes
Very Active Minutes	Lightly Active Minutes
	Total Active Minutes

Polynomial feature derivation is a technique used to transform data by mapping it to a higher dimensional space. This is achieved by increasing the number of subdivisions of the independent variable. The process involves generating new features that capture the interactions and combinations of the original features. When using "PolynomialFeatures" in Sklearn, the original features are transformed by including all possible polynomial combinations up to a specified degree. This expansion effectively creates a higher-dimensional feature space with more subdivisions. For example, if we have two original features X_1 and X_2, and we set the degree parameter to 2, the resulting polynomial features will include not only the original features but also their interactions, such as X_1^2, $X_1 X_2$, and X_2^2. Specifically, two independent variables were chosen from the weak feature subsets, and the highest power was set to 2. The resulting equation is shown in equation (15), and finally, 84 new recombinant features are derived.

$$y = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_1 x_2 + \omega_5 x_2^2 \tag{15}$$

5.1.2. Group Statistical Feature Ferivation

After exploring this correlation between the data, we propose a novel feature engineering method, which divides the values of "Very Active Distance" into different intervals and creates a new feature column named "VD_qcut", as shown in Table 2: When the activity interval ranges from 0km to 1km, the value of VD_qcut is 0, indicating short-distance intense activity; when the activity interval ranges from 1km to 5km, the value of VD_qcut is 1, indicating moderate intense activity; when the activity interval ranges from 5km to 20km, the value of VD_qcut is 2, indicating long-distance intense activity. This method accurately captures the variable correlation patterns between user activity and EE. Furthermore, certain features exhibit ordinal correlations with the user's energy expenditure level. For instance, higher values in the "Total Distance" feature are positively correlated with EE levels, while lower values demonstrate weak correlation. Therefore, using the same method to extract values from different stages of "Total Distance" to create a new ordinal feature column, called "Distance_qcut", with specific interval divisions as shown in Table 2.

Based on the aforementioned binned data, all features were grouped and statistically derived according to the three different value levels (0, 1, 2) in " Very Active Distance " and " Total Distance ". The group-based statistical computations included the mean, maximum, minimum, median, sample variance, and sample total sum. Additionally, the first and third quartiles were also utilized as statistics to obtain derived features with extended statistical metrics.

Table 2. The results of the "Very Active Distance" and "Total Distance" bins.

VD_qcut	Meaning	Distance_qcut	Meaning
Categories		Categories	
0	Daily vigorous exercise distance interval at (0, 1] (measure: km)	0	Daily walking distance interval at (0,4.51] (measure: km)
1	Daily vigorous exercise distance interval at (1, 5] (measure: km)	1	Daily walking distance interval at (4.51,7.387] (measure: km)
2	Daily vigorous exercise distance interval at (5, 20] (measure: km)	2	Daily walking distance interval at (7.387,17.54] (measure: km)

5.1.3. Decision Tree-Based RFECV

The large number of features obtained by the above methods may bring considerable information redundancy. The proposed method integrates Recursive Feature Elimination with Cross-Validation (RFECV) and the decision tree algorithm. RFECV systematically searches for the most relevant feature subset in the feature space by iteratively removing the least relevant feature subsets until the desired number of features is achieved. In this research, RFECV is coupled with the decision tree algorithm, and cross-validation is employed to assess the model's performance on new data, mitigating overfitting and maximizing useful information extraction from limited data. The DF-RFECV algorithm's pseudocode is presented in Table 3.

Table 3. Pseudocode for Decision Tree-Based RFECV Algorithm

Algorithm: Decision Tree-Based RFECV

Input: Decision Tree algorithm, The number of features to be deleted in each round of recursion Step, Cross-validation fold K, Evaluation indicators.

Output: The result array $R[]$ of feature selection.

01: Begin:
02: Divide D into n sample subsets $D_1 \sim D_K$;
03: For $j=1$ to k do:
04: Set the verification set to D_j;
05: Set the training set to $S=\{D_i \mid i=1,2,\cdots,k(i\neq j)\}$;
06: For $i=1$ to m do:
07: S is input Decision-Tree training for training and the importance J of the feature x is calculated;
08: The verification set D_j is input to the trained Decision Tree model to calculate the MSE;
09: Delete the feature $x_{\min}$ corresponding to the minimum feature $J_{\min}$ importance in the training set S and validation set D_j;
11: End For i
12: Save the deleted m feature subsets with the smallest importance;
13: End For j
14: Calculate the $MSE_{average}$ of m feature subsets;

15: Select the feature subset corresponding to the minimum $MSE_{average}$ as the optimal feature subset F;

16: End Begin

5.2. Weight-based Ensemble Learning Voting Strategy

To address the limitations of each individual algorithm, Wei-Voting method is proposed to combine the predictions of multiple ensemble models. Assigning different weights to each model reduces bias to produce more accurate overall predictions. This method aims to improve the model's robustness and enhance its performance by aggregating the average or weighted predictions of the constituent models. Each base learner is assigned a specific weight, which determines its influence on the final prediction. Furthermore, the use of multiple models can help to reduce the risk of overfitting. The prediction value $\hat{y}$ that obtains the highest weighted score can be represented by equation 16.

$$\hat{y} = \sum_{i=1}^{T} w_i h_i(x) / \sum_{i=1}^{T} w_i \tag{16}$$

Where w_i denotes the weight of regressor h_i, which is used to measure the contribution of each learner to the final voting result. T represents the number of participating learners in the voting. $h_i(x)$ represents the prediction result of the i-th base learner, where each learner can provide a prediction value. Then, the prediction result with the highest weighted score is selected as the final voting result.

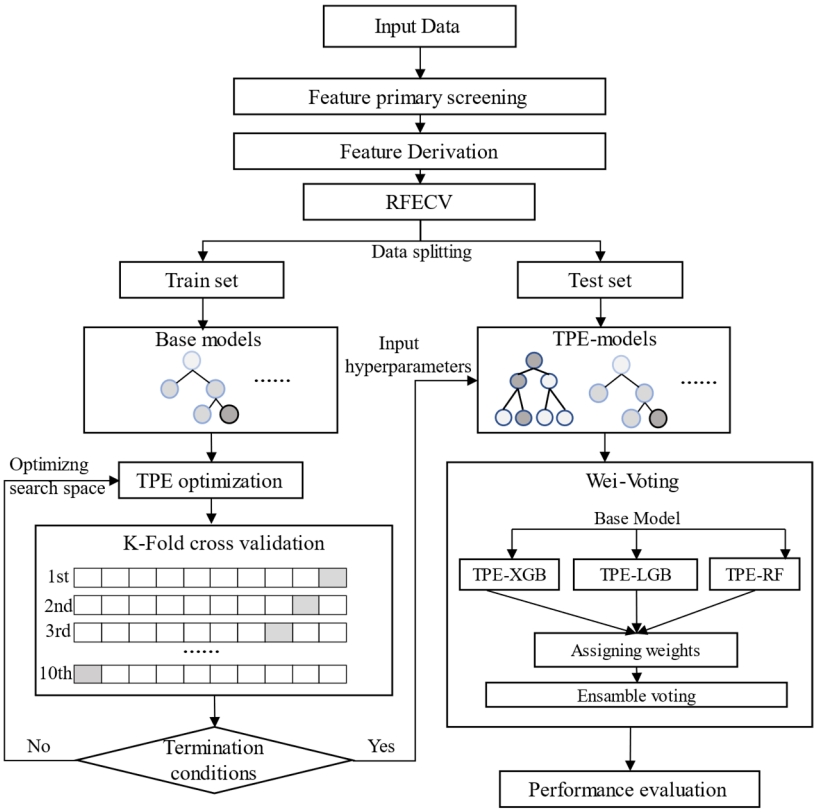

Figure 5. Flowchart of the proposed Wei-Voting algorithm.

The experiments conducted utilized the Wei-Voting algorithm, which incorporates the tree-structured Parzen estimator (TPE) for hyperparameter optimization. The evaluation of the modeling outcomes involves the utilization of 10-fold cross-validation for all three primary regression quantities and Wei-

Voting, and the experimental flowchart for Wei-Voting is depicted in Figure 5. In each TPE hyperparameter optimization, one round of cross-validation was executed.

This paper uses hyperparameter optimization for each base model. Specifically, in the case of XGBoost and LightGBM, the learning rate, maximum depth, subsample ratio, and regularization parameters are mainly optimized, while for random forests, the number of trees, maximum depth, and minimum sample split is the main hyperparameters considered. The TPE hyperparameter optimization technique is employed to tune the parameters of the underlying regressor.

6. RESULTS AND DISCUSSION

6.1. Evaluation Indicators

The efficiency of the EEE model is analyzed herein, with evaluation metrics consisting of the Coefficient of Determination (R^2), Root Mean Squared Error ($RMSE$), and Mean Absolute Error (MAE).The evaluation indicators used are shown as equation 17-19. Optimal model performance is achieved when R^2 approaches 1, and superior performance is indicated by MAE and $RMSE$ values approaching 0. These metrics provide a reliable means of evaluating the model's performance and determining its relevance to the research question.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - y_i^*)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2} \tag{17}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i^* - y_i\right| \tag{18}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i^* - y_i\right)^2} \tag{19}$$

6.2. Prediction Results and Evaluation Analysis

As shown in Figure 6, the Wei-Voting regression algorithm based on the PC-DF-RFECV feature screening strategy performed optimally after the iterations of weight assignment. The R^2 value reached 0.9985, Compared to the R^2 of the Energy Expenditure Regression (EER) model proposed by Lin et al. [19] based on neural network construction, showing a performance improvement of 7.37%, indicating an excellent predictive power and a good fit of the proposed Wei-Voting model, the fitting results are shown in Figure 7 and the fit between the real data and the predicted data was excellent. Furthermore, The MAE of the EEE model, controlled within 5% of the target value's average, achieved a small error and the $RMSE$ result is better than the result of the LAB model (1.07) proposed by Matthew's team [20].This demonstrates that our proposed model outperforms existing methods in EEE, indicating its effectiveness in capturing the complex relationship between the selected features and EE. Compared to the use of bioimpedance analysis for body composition determination and a two-year testing period for subjects, data-driven methods offer advantages of non-invasiveness, portability, and real-time capabilities.

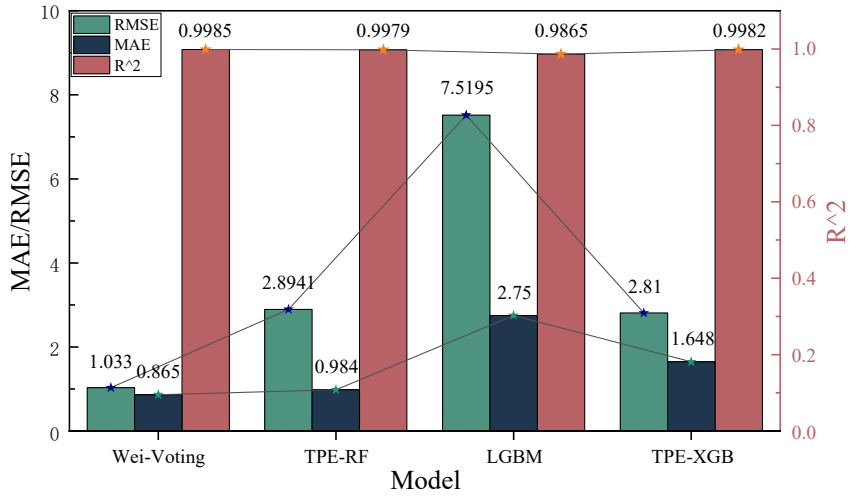

Figure 6. Comparison chart of evaluation metrics of the Wie-Voting algorithm and three base models.

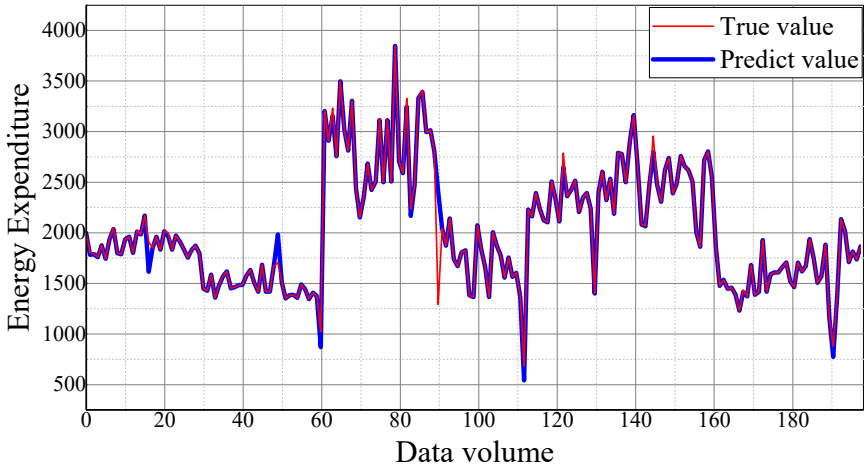

Figure 7. Voting algorithm fitting line graph.

7. CONCLUSION AND FUTURE WORKS

The proposed research is of great significance in the fields of smart homes and activity-based health management-assisted living solutions. By analyzing user data, well-performing machine learning models can provide accurate and personalized health management services, aiding users in assessing exercise intensity, devising workout plans, and improving their overall health condition. Additionally, interdisciplinary collaborations with fields like health sciences, medicine, and engineering are crucial for recommending suitable exercise programs for users and offering activity monitoring and alert functions. The introduction of deep learning and reinforcement learning to enhance model performance is also important in advancing the development of smart home and activity-based health management.

REFERENCES

[1] Parrilla, M.; De Wael, K. Wearable self-powered electrochemical devices for continuous health management. Advanced Functional Materials 2021, 31, 2107042. https://doi.org/10.1002/adfm.202107042

[2] Huifeng, W.; Kadry, S.N.; Raj, E.D. Continuous health monitoring of sportsperson using IoT devices based wearable technology. Computer Communications 2020, 160, 588-595. https://doi.org/10.1016/j.comcom.2020.04.025

[3] Tarafdar, P.; Bose, I. Recognition of human activities for wellness management using a smartphone and a smartwatch: a boosting approach. Decision Support Systems 2021, 140, 113426. https://doi.org/10.1016/j.dss.2020.113426

[4] Shin, J.C.; Kim, J.; Grigsby-Toussaint, D. Mobile phone interventions for sleep disorders and sleep quality: systematic review. JMIR mHealth and uHealth 2017, 5, e7244. https://doi.org/10.1016/j.jad.2022.02.008

[5] Oyelade, T.; Canciani, G.; Carbone, G.; Alqahtani, J.S.; Moore, K.; Mani, A.R. Heart rate variability in patients with cirrhosis: a systematic review and meta-analysis. Physiological Measurement 2021, 42, 055003. DOI 10.1088/1361-6579/abf888

[6] Chowdhury, E.A.; Western, M.J.; Nightingale, T.E.; Peacock, O.J.; Thompson, D. Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. PloS one 2017, 12, e0171720. https://doi.org/10.1371/journal.pone.0171720

[7] Dooley, E.E.; Golaszewski, N.M.; Bartholomew, J.B. Estimating accuracy at exercise intensities: a comparative study of self-monitoring heart rate and physical activity wearable devices. JMIR mHealth and uHealth 2017, 5, e7043. doi:10.2196/mhealth.7043

[8] Wang, Y.; Cang, S.; Yu, H. A survey on wearable sensor modality centred human activity recognition in health care. Expert Systems with Applications 2019, 137, 167-190. https://doi.org/10.1016/j.eswa.2019.04.057

[9] Guo, J.; Yang, L.; Bie, R.; Yu, J.; Gao, Y.; Shen, Y.; Kos, A. An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring. Computer Networks 2019, 151, 166-180. https://doi.org/10.1016/j.comnet.2019.01.026

[10] Sattar, S.; Li, S.; Chapman, M. Road surface monitoring using smartphone sensors: A review. Sensors 2018, 18, 3845. https://doi.org/10.3390/s18113845

[11] Yuan, Y.; Shen, Q.; Wang, S.; Ren, J.; Yang, D.; Yang, Q.; Fan, J.; Mu, X. Coronavirus mask protection algorithm: A new bio-inspired optimization algorithm and its applications. Journal of Bionic Engineering 2023, 1-19.

[12] Yang, D.-H.; Lee, H.-J.; Lim, D.-J. RolexBoost: A Rotation-Based Boosting Algorithm With Adaptive Loss Functions. IEEE Access 2020, 8, 41037-41044. DOI: 10.1109/ACCESS.2020.2976822

[13] Ren, J.; Wang, Z.; Pang, Y.; Yuan, Y. Genetic algorithm-assisted an improved AdaBoost double-layer for oil temperature prediction of TBM. Advanced Engineering Informatics 2022, 52, 101563. https://doi.org/10.1016/j.aei.2022.101563

[14] Zhou, Z.-H.; Wu, J.; Tang, W. Ensembling neural networks: many could be better than all. Artificial intelligence 2002, 137, 239-263. https://doi.org/10.1016/S0004-3702(02)00190-X

[15] Kim, H.; Kim, H.; Moon, H.; Ahn, H. A weight-adjusted voting algorithm for ensembles of classifiers. Journal of the Korean Statistical Society 2011, 40, 437-449. https://doi.org/10.1016/j.jkss.2011.03.002

[16] van den Heuvel, E.; Zhan, Z. Myths about linear and monotonic associations: Pearson'sr, Spearman's ρ, and Kendall's τ. The American Statistician 2022, 76, 44-52. https://doi.org/10.1080/00031305.2021.2004922

[17] Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T. Xgboost: extreme gradient boosting. R package version 0.4-2 2015, 1, 1-4.

[18] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 2017, 30.

[19] Lin, C.-W.; Yang, Y.-T.C.; Wang, J.-S.; Yang, Y.-C. A wearable sensor module with a neural-network-based activity classification algorithm for daily energy expenditure estimation. IEEE Transactions on Information Technology in Biomedicine 2012, 16, 991-998. 10.1109/TITB.2012.2206602

[20] Hibbert, J.M.; Broemeling, L.D.; Isenberg, J.N.; Wove, R.R. Determinants of Free-Living Energy Expenditure in Normal Weight and Obese Women Measured by Doubly Labeled Water. Obesity research 1994, 2, 44-53. https://doi.org/10.1002/j.1550-8528.1994.tb00043.x