CalFit: Physiology-Aware Gradient Boosting with Conformal Intervals for Reliable Calorie Prediction

Shaik Haseeb Ur Rahman
Department of Computer Science
University of California, Davis
One Shields Ave, Davis, CA 95616
hrahman@ucdavis.edu

Preyash Yadav
Department of Computer Science
University of California, Davis
One Shields Ave, Davis, CA 95616
preyadav@ucdavis.edu

Nicolai Amann, Ph.D.*
Department of Statistics
University of California, Davis
One Shields Ave, Davis, CA 95616
ndamann@ucdavis.edu

Team Membership and Attestation

All team members listed below consent to the content of this report, affirm the originality of this work, and commit to upholding academic integrity throughout the project. Each member's contribution is mentioned below:

- ***Shaik Haseeb Ur Rahman***: Contributed to the Literature Review, Background Work, Research Question Formulation, Data Acquisition, Data Pre-processing, Exploratory Data Analysis, end-to-end RQ#1 and RQ#3 Analysis, and finally post-discussion analysis.
- ***Preyash Yadav***: Contributed to the Literature Review and Background Work, Research Question formulation, Data Acquisition, Data Pre-processing, Exploratory Data Analysis, end-to-end RQ#1 and RQ#2 Analysis, and finally post-discussion analysis.

Abstract

Accurate estimation of energy expenditure or calories burnt during exercise is central and necessary for health monitoring, fitness planning, and weight-management goals. Yet popular calorie estimators and commercial fitness trackers frequently rely on generic formulas that underutilize rich physiological data and fail to account for inter-individual differences in various factors like age, gender, height, weight, heart rate, duration, body temperature, and exercise intensity, leading to biased and unreliable outputs. CalFit addresses this gap by formulating calorie prediction as a supervised learning problem over a Kaggle-sourced dataset of workout sessions comprising physiological, anthropometric, demographic, and session-specific features. Methodologically, we benchmark a transparent LASSO regression model against a monotone gradient-boosted decision tree (GBDT) that is constrained to produce non-decreasing calorie predictions with respect to key physiological drivers, particularly duration and heart rate. This design explicitly encodes basic domain knowledge to avoid implausible model behavior. To move beyond point predictions, we wrap the selected best model with split-conformal prediction, yielding distribution-free 95% prediction intervals that provide honest, finite-sample uncertainty quantification at both the overall and subgroup levels. We evaluate performance under an 80/20 train–test split with 5-fold cross-validation on the training set, reporting RMSE, MAE, and MAPE, as well as coverage and mean interval width. It was observed that the Monotone GBDT model outperformed the other models with an RMSE of 2.526 and R^2 of 0.998. CalFit aims to deliver calorie estimates that are not only accurate, but also physiologically plausible and uncertainty-aware, increasing their usefulness for real-world decision making.

Keywords: ***LASSO Regression, Monotone-GBDT, Partial Dependence Plot, Individual Conditional Expectation and Split-conformal Prediction***

*Project CalFit, completed by Shaik Haseeb Ur Rahman (924142853) and Preyash Yadav (924168918), was carried out under the guidance of Nicolai Amann, Ph.D., for STA 221: Big Data & High-Dimensional Statistical Computing.

1 Introduction / Motivation

Calories burned estimations are widely used to guide workout planning, monitor progress toward fitness and weight-management goals, and provide feedback within calorie estimators, digital health platforms, and wearable devices like Apple Watch and Fitbit. In practice, these estimations influence user decisions about training intensity, workout duration, recovery, and dietary adjustments. Consequently, system errors in calorie estimation can propagate into downstream decisions, potentially leading to misleading feedback.

Despite their prevalence, many real-world calorie estimators are grounded in simplified modeling assumptions. They frequently rely on demographic inputs like age, gender, and weight, and generic relationships between these factors and energy expenditure, without fully leveraging physiological factors observed during exercise, like heart rate dynamics or body temperature. This simplification can yield overly generalized predictions that fail to reflect meaningful variability across individuals and sessions. In addition, purely data-driven machine learning models, while often improving predictive accuracy, can introduce two further limitations that are particularly consequential in health-related contexts.

First, many high-performing non-linear models behave as black boxes and do not explicitly enforce basic physiological plausibility. A flexible model also may occasionally produce predictions that contradict fundamental expectations, like lower predicted calories for a higher heart rate or longer duration. Even though such violations occur infrequently, they can undermine user trust, complicate scientific interpretation, and raise concerns about the reliability of model behavior under distribution shift or in underrepresented subgroups.

Second, prior work commonly emphasizes point prediction accuracy while neglecting uncertainty quantification. A single numerical estimate provides no indication of reliability for a given workout, which is problematic when measurement noise, limited sample support, or subgroup imbalance can materially affect predictive confidence.

CalFit is motivated by the need for a more trustworthy modeling pipeline that balances three objectives:

- (i) Rigorous comparison between interpretable linear baselines and flexible non-linear models,
- (ii) Physiologically grounded constraints that reduce the risk of implausible predictions, and
- (iii) Calibrated uncertainty quantification that communicates reliability at the individual-session level

To achieve this, CalFit compares LASSO regression against a monotone GBDT model that enforces non-decreasing dependence on duration and heart rate, and then wraps the chosen model with split-conformal prediction to yield distribution-free prediction intervals with approximately nominal coverage. This combined approach is intended to improve the reliability and real-world usability of calorie-burn prediction.

2 Background

2.1 Problem definition

Let each workout session be represented by a feature vector $x \in \mathbb{R}^p$ that includes physiological, anthropometric, demographic, and session-level predictors. In CalFit, these features include heart rate, body temperature, height, weight, BMI, age, gender (encoded appropriately), and duration. Let the target variable be $y \in \mathbb{R}$, the total calories burned during the session. The primary objective is to learn a predictive function that outputs a point estimate $\hat{y} = f(x)$ of calories burned for a given session.

$$f : \mathbb{R}^p \to \mathbb{R}$$

In addition to minimizing predictive error on unseen data, CalFit imposes a ***structural plausibility*** requirement on the learned mapping. Specifically, for key physiological drivers x_{dur} (duration) and x_{hr} (heart rate), the model should satisfy monotonicity constraints of the form:

$$\frac{\partial f(x)}{\partial x_{\text{dur}}} \geq 0, \qquad \frac{\partial f(x)}{\partial x_{\text{hr}}} \geq 0,$$

These constraints encode the domain expectation that, all else equal, *increasing workout duration or heart rate should not decrease the predicted total calories burned.* The monotone GBDT model is used to operationalize these constraints while preserving non-linear capacity to capture interactions and heterogeneity across individuals and sessions.

Beyond point prediction, CalFit also defines a second objective: to produce an uncertainty quantification mechanism that returns a prediction interval $[L(x), U(x)]$ such that, for a new session (X, Y),

$$\mathbb{P}\big(Y \in [L(X), U(X)]\big) \approx 0.95,$$

under minimal assumptions about the data distribution. This is achieved via split-conformal prediction, which leverages a calibration set to convert model residuals into distribution-free intervals with finite-sample marginal coverage guarantees.

2.2 Prior Work

Nipas et al. [1] used a dataset from Kaggle with 15000 samples with following features: body temperature, gender, age, weight/height (BMI), heart rate, The research work used regression algorithms like ridge regression, linear regression, and random forest to build a calorie-burn prediction model, with the Random Forest achieving the best results with 95.77% accuracy and 2.85 RMSE. The paper made assumptions that the dataset's variables/features are enough to predict the calorie burn. It also assumed that the physiological factors are uniform regardless of the fitness level or metabolism of different users. The limitations in this paper were to just rely on basic regressions, without exploring any ensemble approach. The study had no personalization based on the user's metabolism differences or fitness levels.

This research work in Kadam et al. [2] is a buildup on Nipas et al. They performed data preprocessing like standardization, encoding, and basic hyperparameter tuning like GridSearchCV. The model performance was evaluated using MAE, MSE, and RMSE. They assumed that calorie burn only relies on the physiological and demographic features and that RMSE is a sufficient measure of performance. The research gap is that they also just used a Random Forest approach; the high value of RMSE might indicate possible overfits. The research also suggested using more data points over time to test model performance.

Prasad et al. [3] utilize regression models like linear regression, ridge regression, XGBoost regression, etc., and perform a comparative study for calorie-burn prediction. The study discovered that ensemble models (XGBoost) capture nonlinear patterns better than linear regressors evaluated by MAE, MSE, and R^2. The study assumes that calorie burn doesn't depend upon factors like time series, metabolic rate, medical conditions, etc. The research gap falls on the way they collected the dataset from GeeksForGeeks, which is not a very reliable source.

Reddy et al. [4] performed a comparative study on XGBoost and a novel Random Forest Regressor, each reaching a model accuracy of 83.73% and 72.26%, for burn-calorie predictions. The research utilized SPSS and an independent sample t-test for statistical analysis and to compare model accuracies. The study assumed that statistical tests like the t-test are appropriate measures for model differences. The research gap here is having a very small dataset with only 200 samples, which limits the model generalization capabilities, and it only focuses on overall prediction accuracy rather than finding relationships between features, etc.

2.3 Need of this study

As we learned in the prior works, they mainly focus on point accuracy and often deploy flexible regression models, but the main issues in them are:

- Comparability across model classes is unclear without a principled linear baseline
- Plausibility is not enforced (no monotone constraints), risking nonsense extrapolations
- Uncertainty is not calibrated: users see a distinct label with unknown reliability

Filling these gaps increases the trust and actionability of calorie predictions and hence leads to a reliable prediction.

2.4 Research Questions

To cover the gaps of the prior works, our research is guided by 3 key questions:

- **RQ#1:** Between LASSO and Monotone GBDT, which generalizes better under 5-fold CV?
- **RQ#2:** Does Monotone GBDT outperform LASSO while showing non-decreasing Partial Dependence/Individual Conditional Expectation for duration and heart rate?

- **RQ#3:** Do split-conformal intervals achieve 95% coverage overall and across male/female and age bins, with reasonable width?

3 Methodology

3.1 Dataset

We used a Kaggle-sourced Calorie Burnt dataset, which contains **15,000 workout sessions** data with typical physiological vitals, duration, and calories burnt label. The primary target variable (output label) is total calories burned per session. For reproducibility, we create a single 80/20 train–test split with a fixed random seed and perform 5-fold cross-validation within the training set for model selection and tuning. If a user_id is present, we employ GroupKFold so that all sessions from the same individual stay within the same fold and never appear across train/validation/test, preventing leakage and overly optimistic performance estimates.

3.2 Features/Metrics of Interest

The dataset features include:

- **Physiological Features:** heart rate, body temperature
- **Anthropometric Features:** height, weight, BMI
- **Demographic Features:** age, gender
- **Session-specific Features:** duration

Target Variable: Total calories burnt per session
Error Metrics: RMSE and MAE
Uncertainty Metrics: Coverage of 95% intervals and mean interval width, overall and by gender and age bins

3.3 Data Pre-processing

In the data preprocessing stage, we first loaded the raw workout-session dataset and standardized it into a clean, analysis-ready pandas DataFrame. We validated the schema and datatypes, removed duplicate or invalid records, and handled missing values to ensure model inputs were complete and consistent. Next, we normalized feature naming and performed light feature engineering to improve signal quality. Finally, we prepared the modeling matrix by separating predictors and target, encoding categorical variables (e.g., gender) as numeric, and applying scaling where required (particularly for linear models such as LASSO), producing a reproducible pipeline suitable for cross-validation and fair comparison against tree-based models.

3.4 Exploratory Data Analysis

3.4.1 Distribution of Calories

- ***Right-skewed target:*** Calories burned per session has mean 89.5 and median 79.0, with most values between approximately 35 and 138 calories and a long right tail up to about 314 calories (skew ≈ 0.51). This indicates a concentration of short-to-moderate workouts with relatively few very long or very intense sessions.
- ***Implications for model class:*** Because the response is clearly right-skewed rather than Gaussian, models that assume a symmetric, homoscedastic target may be suboptimal. This motivates the use of flexible, non-linear methods (e.g., tree-based models and monotone GBDT) and the consideration of transformations if needed.
- ***Focus of evaluation:*** Since most users lie in the low–to–mid calorie range, our evaluation of predictive performance and interval coverage emphasizes this dense region of the distribution, while still checking that the model behaves sensibly in the high-calorie tail.

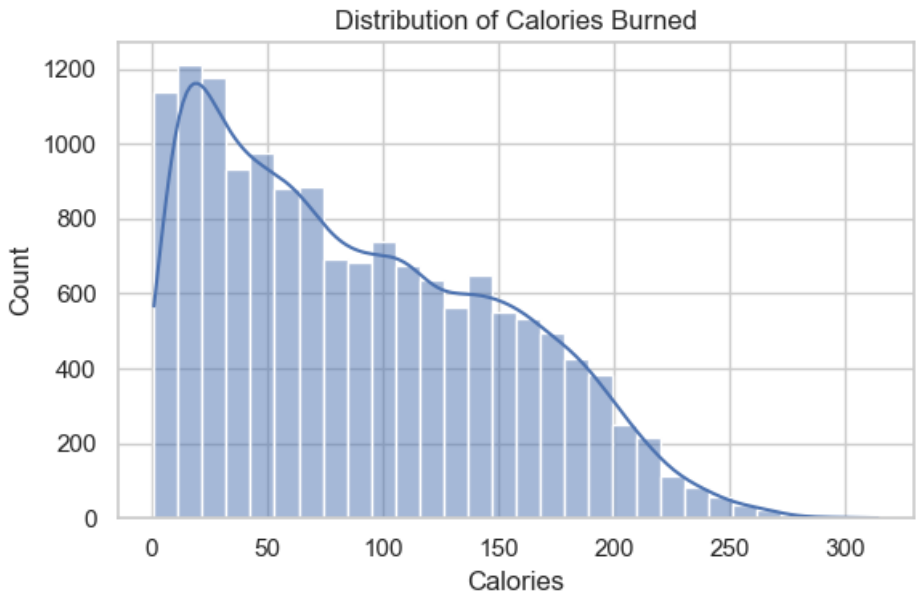


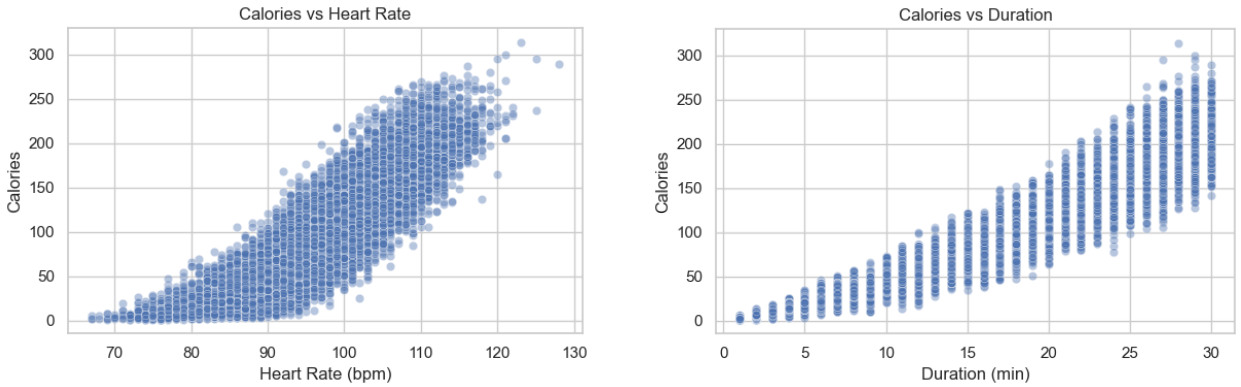
Figure 1: Distribution of calories burned per workout session.

3.4.2 Calories v/s Heart Rate and Duration

- ***Duration as a near-linear driver:*** Calories exhibit a very strong, almost linear increase with Duration (corr ≈ 0.96), with relatively modest spread at each duration value. This confirms Duration as a primary determinant of energy expenditure and supports enforcing a monotonic constraint "Duration $\uparrow \Rightarrow$ Calories $\uparrow$" in monotone GBDT.

- ***Heart rate as an intensity proxy:*** Calories also rise sharply with Heart_Rate (corr ≈ 0.90), though with more vertical variability than for Duration. This pattern reflects that heart rate captures workout intensity: sessions of similar length can burn substantially different calories depending on how hard the participant is working.

- ***Joint modeling motivation:*** Together, these plots justify modeling Calories as a function of both Duration and Heart_Rate to distinguish, for example, a long low-intensity session from a shorter high-intensity one. This motivates our choice of models that can handle interactions and non-linearities (e.g., monotone GBDT) and underpins the monotonicity constraints used in our research questions.



(a) Calories vs Heart Rate

(b) Calories vs Duration

Figure 2: Relationship of Calories with Heart Rate and Duration.

3.4.3 Correlation Heatmap

- ***Primary drivers of calories:*** The strongest positive correlations with Calories are Duration ($r \approx 0.96$), Heart_Rate ($r \approx 0.90$), and Body_Temp ($r \approx 0.82$). Weight shows a weaker but directionally positive relationship, while Age, Gender, Height, and BMI are only weakly correlated with the target.

- ***Feature selection and interpretation:*** These correlations identify Duration, Heart_Rate, Body_Temp, and (to a lesser extent) Weight as the main drivers our models should prioritize. Demographic variables play a secondary role once actual workout behavior is observed, which helps us interpret variable-importance patterns later in the modeling stage.

- ***Modeling constraints and multicollinearity:*** The strong correlations among Height, Weight, and BMI highlight potential multicollinearity issues for linear models such as LASSO, suggesting the need for regularization and careful interpretation of coefficients. For tree-based or GBDT models, the heatmap guides where to impose monotonicity (Duration, Heart_Rate, Body_Temp) and informs our expectations when analyzing partial dependence and ICE plots in the research questions.

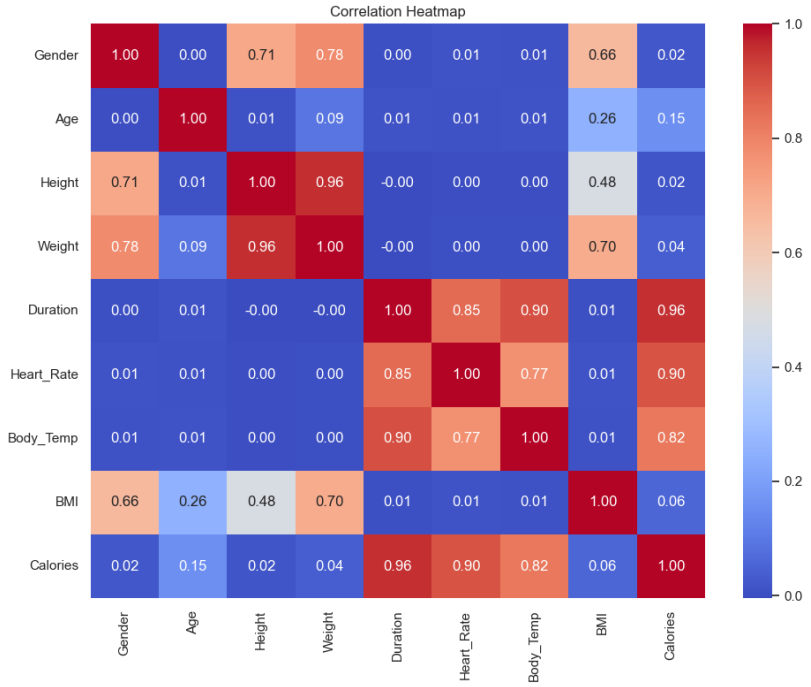


Figure 3: Correlation matrix for physiological, anthropometric, and session-level features.

3.5 Methodology

3.5.1 Algorithms Used

- **LASSO Regression:** LASSO (Least Absolute Shrinkage and Selection Operator) is a linear regression method that performs coefficient shrinkage and variable selection via an ℓ_1 penalty. Given data $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, LASSO estimates $(\beta_0, \beta) \in \mathbb{R} \times \mathbb{R}^p$ by

$$(\hat{\beta}_0, \hat{\beta}) \in \arg\min_{\beta_0, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i^\top \beta\right)^2 + \lambda \|\beta\|_1 \right\},$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ and $\lambda \geq 0$ controls the amount of regularization. Larger λ yields sparser $\hat{\beta}$ by shrinking some coefficients exactly to zero.

- **Monotone Gradient-Boosted Decision Trees:** Gradient-boosted decision trees build an additive model

$$f_M(x) = \sum_{m=1}^{M} \nu\, h_m(x),$$

where each h_m is a regression tree fit to the negative gradient of a chosen loss (e.g., squared error) and $\nu \in (0, 1]$ is a learning rate. In a *monotone* GBDT, for a subset of features $\mathcal{S}$ we constrain f_M to be monotone in those coordinates. For a feature $j \in \mathcal{S}$ with desired non-decreasing behavior, the model is constrained to satisfy

$$x'_j \geq x_j \;\Rightarrow\; f_M(x_1, \ldots, x'_j, \ldots, x_p) \;\geq\; f_M(x_1, \ldots, x_j, \ldots, x_p),$$

for all relevant feature vectors x. These constraints are enforced during tree construction (e.g., by restricting split directions and leaf values) so the final ensemble respects known domain monotonicities (such as "duration $\uparrow \Rightarrow$ calories $\uparrow$").

- **Split-Conformal Prediction:** Split-conformal prediction provides distribution-free prediction intervals with finite-sample coverage under exchangeability. Given data $\{(x_i, y_i)\}_{i=1}^{n}$:

  1. Randomly split indices into a proper training set $\mathcal{I}_{\text{train}}$ and a calibration set $\mathcal{I}_{\text{cal}}$.
  2. Fit any point predictor $\hat{f}$ using only $\mathcal{I}_{\text{train}}$.
  3. Compute calibration residuals
  $$r_i = \big|y_i - \hat{f}(x_i)\big|, \qquad i \in \mathcal{I}_{\text{cal}}.$$
  4. For target miscoverage α (e.g., $\alpha = 0.05$), let $q_{1-\alpha}$ be the empirical $(1-\alpha)$-quantile of $\{r_i\}_{i \in \mathcal{I}_{\text{cal}}}$ with the usual conformal correction. The $(1-\alpha)$ prediction interval for a new input x_{new} is
  $$C(x_{\text{new}}) = \Big[\hat{f}(x_{\text{new}}) - q_{1-\alpha},\; \hat{f}(x_{\text{new}}) + q_{1-\alpha}\Big].$$

  Under exchangeability of (X_i, Y_i) and $(X_{\text{new}}, Y_{\text{new}})$, split-conformal prediction guarantees $\mathbb{P}\big(Y_{\text{new}} \in C(X_{\text{new}})\big) \geq 1 - \alpha$ in finite samples, regardless of the underlying distribution or the specific choice of $\hat{f}$.

- **Polynomial LASSO (LASSO-Poly):** Apply a polynomial feature expansion $\phi(\mathbf{x})$ (e.g., all monomials up to degree d) and fit a linear model with an ℓ_1 penalty to induce sparsity:

$$\hat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \phi(\mathbf{x}_i)^\top \boldsymbol{\beta}\Big)^2 \;+\; \lambda \|\boldsymbol{\beta}\|_1 \right\},$$

where $\lambda \geq 0$ controls regularization strength and $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$. (Typically β_0 is not penalized.)

3.5.2 Evaluation Methods

- **Root Mean Squared Error (RMSE):** RMSE measures the typical size of prediction errors, with large errors penalized more heavily. Given observed targets y_i and predictions $\hat{y}_i$ for $i = 1, \ldots, n$,

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \big(y_i - \hat{y}_i\big)^2}.$$

It is expressed in the same units as the response (here: calories) and is sensitive to outliers.

- **Mean Absolute Error (MAE):** MAE is the average absolute deviation between predictions and observations:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \big|y_i - \hat{y}_i\big|.$$

Compared to RMSE, MAE weights all errors linearly and is therefore more robust to large outliers.

- **Partial Dependence Plot (PDP):** PDPs summarize the *marginal effect* of one feature (or a small set of features) on the model prediction by averaging over the empirical distribution of all other features. For a single feature x_j and a

fitted model f, the partial dependence function is

$$\mathrm{PD}_j(z) = \frac{1}{n}\sum_{i=1}^{n} f\big(z, x_{i,-j}\big),$$

where $x_{i,-j}$ denotes all components of x_i except x_{ij}. Plotting $\mathrm{PD}_j(z)$ against z reveals how the model's prediction changes as x_j varies, averaged over the data.

- **Individual Conditional Expectation (ICE):** ICE curves show the *instance-specific* effect of a feature, by varying one feature for a fixed observation while holding all others constant. For feature x_j, the ICE curve for observation i is

$$\mathrm{ICE}_{i,j}(z) = f\big(z, x_{i,-j}\big).$$

  Plotting $\mathrm{ICE}_{i,j}(z)$ for many i reveals heterogeneity in feature effects that may be hidden by the averaged PDP, and helps diagnose interactions and non-linearities.

- **Coefficient of Determination (R^2), Learning Curves, and Permutation Importance:** The coefficient of determination R^2 quantifies the fraction of variance in y explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\big(y_i - \hat{y}_i\big)^2}{\sum_{i=1}^{n}\big(y_i - \bar{y}\big)^2},$$

  where $\bar{y}$ is the sample mean of y_i. We examine *learning curves* of R^2 as a function of training set size to diagnose underfitting vs. overfitting (e.g., whether additional data would still improve generalization). We also compute *permutation importance* by randomly permuting feature x_j in the test set, recomputing R^2, and measuring the drop ΔR^2_j; large drops indicate features that are crucial for predictive performance.

3.5.3 Assumptions Adopted

- **Feature and Label Reliability:** The input features (e.g., heart rate, duration, body temperature) and the target *Calories* are assumed to be measured consistently and accurately enough that the learned relationships reflect physiology rather than noise or systematic bias.
- **I.I.D. / Exchangeability:** Each observation is assumed independent and drawn from the same underlying distribution. This is required for split-conformal prediction intervals to have valid marginal coverage.
- **No Distribution Shift:** The train, calibration, and test splits are assumed to come from the same data-generating process, so model performance and conformal coverage generalize.

3.5.4 Methodological Innovations

- **Physiology-aligned modeling via Monotone GBDT:** We imposed monotonic constraints on physiologically meaningful features (e.g., *Duration* and *Heart Rate*) so the learned mapping respects expected relationships (increasing effort should not decrease predicted calories), improving interpretability and plausibility.
- **Uncertainty quantification via Split Conformal Prediction:** Beyond point predictions, we produced calibrated prediction intervals using a separate calibration split, enabling distribution-free uncertainty estimates for more reliable, risk-aware calorie predictions.

4 Result Analysis

This section describes the results corresponding to each research questions (RQ1-RQ3).

4.1 RQ#1: Between LASSO and Monotone GBDT, which generalizes better under 5-fold CV?

To evaluate which modeling approach generalizes better, both LASSO and Monotone GBDT were compared using 5-fold cross-validation on the training sets. To evaluate the performances, the Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) were used as the metrics.

It was observed that LASSO achieved a mean RMSE of 11.033 and a mean R^2 of 0.969 while the Monotone GBDT achieved a much lower RMSE of 2.526 and a higher R^2 value as presented in Table 1

Table 1: 5-fold cross-validation results

Model	RMSE (mean)	RMSE (std)	R^2 (mean)	R^2 (std)
LASSO	11.033	0.168	0.969	0.001
Monotone GBDT	2.526	0.090	0.998	0.000

The consistent lower RMSE and reduced variance values of Monotone GBDT across folds indicate that Monotone GBDT generalizes better than LASSO under cross-validation. These results suggest that the Monotone GBDT performs a better fit than LASSO for the calorie burn prediction under cross-validation.

4.2 RQ#2: Does Monotone GBDT outperform LASSO while showing non-decreasing Partial Dependence/Individual Conditional Expectation for duration and heart rate?

To compare the predictions of LASSO and Monotone GBDT, both the models were trained on training split and evaluated on an unseen test-split. The model performance was then evaluated using metrics like the RMSE, MAE, and R^2.

On the test set, LASSO achieved an RMSE of 11.224, MAE of 8.193 and R^2 of 0.969. Monotone GBDT achieved an RMSE of 2.362, MAE of 1.660 and r^2 of 0.999 which shows better performance than LASSO. The significant drop in both the RMSE and MAE values indicate that Monotone GBDT provides better prediction on calorie burn data than LASSO on unseen data.

We also examined monotonicity with respect to few exercise-related features. Using the trained Monotone GBDT model, Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) curves were generated for exercise duration and heart rate. The PDP and ICE curves exhibited a non-decreasing behavior for both the features as evident in Figure 4. A numerical monotonicity check further confirmed that the predicted calorie burn did not decrease with increase in duration or heart rate increase.

The results demonstrate that Monotone GBDT outperforms LASSO in terms of calorie burn prediction and also produces predictions that are physiologically expected in monotonic relationships between calorie burn, exercise duration and heart rate.

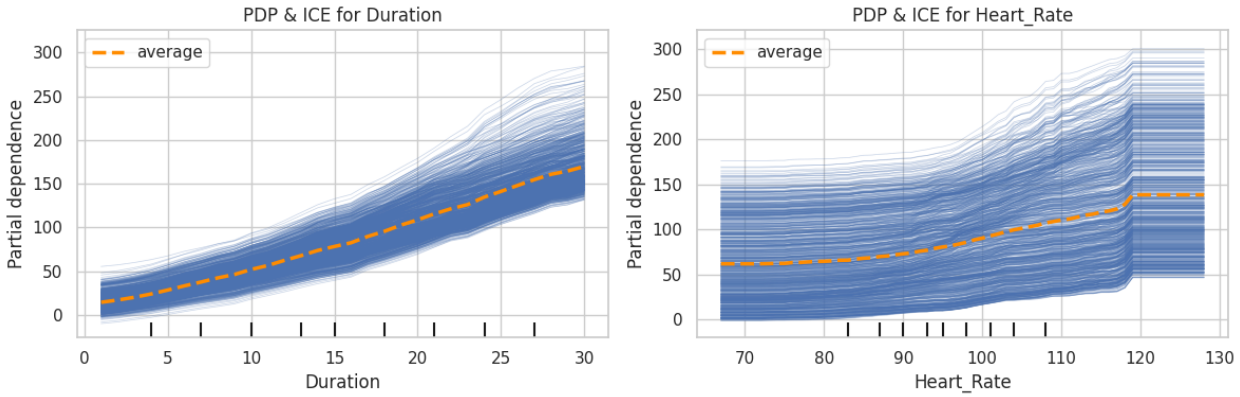


Figure 4: PDP & ICE for Monotone GBDT.

Table 2: Performance Comparison

Model	RMSE	MAE	R^2
LASSO	11.224	8.193	0.969
Monotone GBDT	2.362	1.660	0.999

4.3 RQ#3: Do split-conformal intervals achieve 95% coverage overall and across male/female and age bins, with reasonable width?

To assess the uncertainty in the predictions, we utilized split-conformal intervals using a separate calibration set. To reach 95% coverage the conformal quantile was estimated to be 5.094 calories.

When evaluation was performed on an unseen test set, the prediction interval covered the true value in 95.93% cases. This value is very close to our desired 95% coverage level. The average width of the prediction intervals was 10.188 calories which indicates a reasonably narrow uncertainty bounds. The coverage for different demographic groups were also analyzed and the results can be seen in Table 3 and Table 4.

Table 3: Coverage by Gender Bin

Gender	Coverage	Avg_Width	n
0	97.62%	10.188	1510
1	94.23%	10.188	1490

Table 4: Coverage by Age Bin

Age Bin	Coverage	Avg_Width	n
$\leq$25	97.80%	10.188	546
26–40	96.74%	10.188	983
41–60	95.72%	10.188	865
60+	93.23%	10.188	606

These results show that for the full test set and the different gender, age groups the split-conformal intervals does provide well-calibrated and reliable estimates of uncertainty.

5 Discussion

This section addresses the fundamental questions encountered during our presentation on November 25, 2025. In this section, we will be implementing the approaches recommended by Prof. Amann and our peers. We will be analyzing the results and comparing them with our previous approaches, in order to improve the quality of our study.

5.1 What happens when we try Poly LASSO?

As we were advised to try adding polynomial features in LASSO to assess, if it performs better and captures the non-linear relationship better. We trained LASSO model with a second-order polynomial features - squared heart rate, squared duration, and an interaction between heart rate and duration. We observed that this approach improved the performance of prediction as compared to the Linear LASSO model, achieving a lower RMSE and higher R^2 on the test set. These results also indicate that calorie burn in impacted by factors like exercise duration and heart rate in a non-linear way.

Although the polynomial LASSO performed better than the linear LASSO, it was observed that the polynomial LASSO doesn't guarantee the model to suit the physiological expectations. The Figure 5 shows that the Polynomial LASSO predicted fewer calories burn for a longer duration, which is **physiologically unrealistic**. In comparison to this, Monotone GBDT, as shown in Figure 4, clearly aligns well with the physiological expectations.

To analyze it further, a strict monotonicity check was performed by increasing the workout duration while keeping the other features fixed. It was observed that the polynomial LASSO model violated monotonicity in a large number of cases; in comparison, the Monotone GBDT didn't perform any such violation.

These results indicate that although adding non-linearity in linear models does improve the model performance, it doesn't guarantee the correct or expected physiological behavior, hence indicating us to prefer gradient-boosted models rather than regression models for this kind of problem statement.

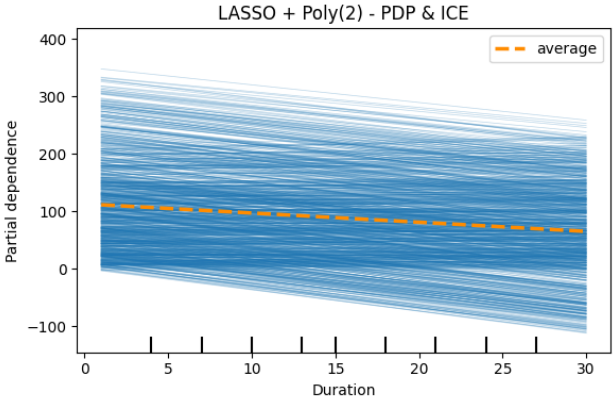
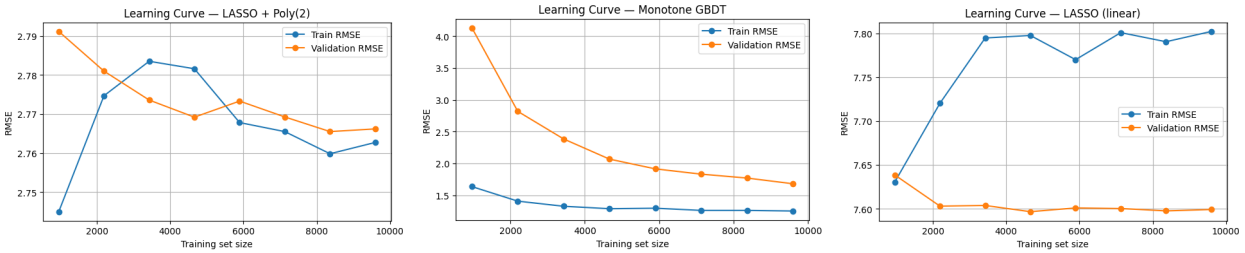


Figure 5: PDP & ICE for Polynomial LASSO

5.2 Are our models overfitting?

We observed that the models achieved the R^2 values close to 1, which raises concerns on model overfitting. However, in this dataset, our target variable calorie burn is highly correlated with exercise duration and heart rate, as evident in the Figure 3, hence having a high R^2 is not sufficient evidence to analyze overfitting. To assess this, we performed some additional steps such as the Learning Curve Analysis and Permutation Tests.

Learning Curve Analysis: For this analysis we plotted the training and validation RMSE against the training size. Figure 6 shows that for the models LASSO(Polynomial) and Monotone GBST, the validation error consistently dropped and approached the training error. It is also evident that the gap between training and validation error is very less as the training size increases which indicates that the model is generalizing well on this dataset and is not memorizing the training data. But for LASSO (Linear), the training and validation errors had an early convergence, which suggests underfitting rather than overfitting. Refer to the Learning Curve Tables in Appendix A for more details.



(a) Learning Curve for LASSO (Poly) (b) Learning Curve for Monotone GBDT) (c) Learning Curve for LASSO (Linear)

Figure 6: Learning Curves

Permutation Tests: For this analysis, we shuffled the key features like duration and heart rate randomly. From Table 5, Table 6 we can observe that permuting the duration caused a very significant degradation in model performance for both LASSO(Polynomial) and Monotone GBDT, with some RMSE values increasing significantly while some R^2 values becoming negative. Permuting the heart rate also had significant effects on the performance of non-linear models but for linear LASSO as evident in the Table 7 it was minimal, which also suggests us that linear models are not suited for this use case.

Table 5: Permutation Test Results: LASSO Polynomial

Feature	base_rmse	perm_rmse_mean	rmse_increase	base_r2	perm_r2_mean	r2_drop
Duration	2.788596	65.130713	62.342123	0.998073	-0.051255	1.049329
Heart_Rate	2.788593	31.804811	29.016221	0.998073	0.749306	0.248767

Table 6: Permutation Test Results: Monotone GBDT

Feature	base_rmse	perm_rmse_mean	rmse_increase	base_r2	perm_r2_mean	r2_drop
Duration	1.627333	65.068388	63.441055	0.999344	-0.049240	1.048584
Heart_Rate	1.627333	30.033100	28.405767	0.999344	0.776469	0.222874

Table 7: Permutation Test Results: LASSO (Linear)

Feature	base_rmse	perm_rmse_mean	rmse_increase	base_r2	perm_r2_mean	r2_drop
Duration	7.916542	80.768175	72.851633	0.984471	-0.616534	1.601005
Heart_Rate	7.916542	7.944718	0.028176	0.984471	0.984360	0.000111

The results of these analysis indicate that the high R^2 values observed in our models are not due to overfitting, the models learn the strong relationships between duration, heart rate, and calorie burn, and their performance proves to be stable in our analysis.

5.3 Future Work

For future work, we plan to extend CalFit as a scalable, user facing application by deploying a Gradio-based UI that accepts workout inputs and returns both a point estimate and an uncertainty-aware calorie range. Concretely, the application backend will be served by the monotone GBDT, and the intermediate output will be wrapped with split-conformal prediction to provide distribution-free prediction intervals with approximately nominal coverage, enabling more reliable decision support in practical fitness settings.

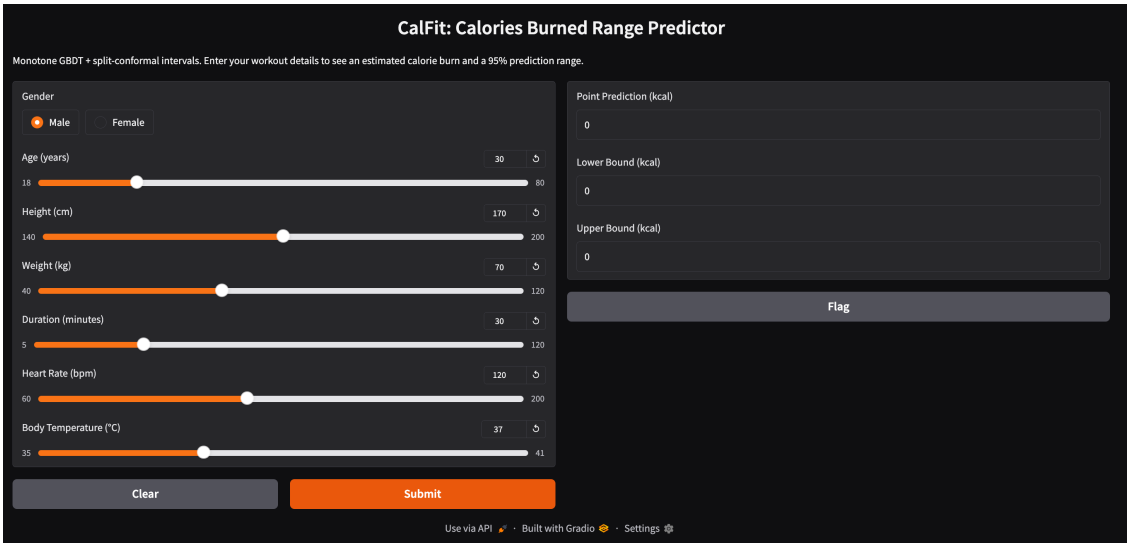


Figure 7: Gradio App

6 Conclusion

In this study, we observed that for problems involving physiological factors such as heart rate, and exercise duration, linear models (like LASSO) are often insufficient in learning the underlying relationship between the factors. Although, we observed that by adding a polynomial feature in LASSO, the performance of this model improved, but the predictions results were not physiologically plausible. As polynomial LASSO model produced results that show a decrease in predicted calorie value as the duration of exercise increases which is unrealistic. On the contrary, the monotone GBDT model managed to learn the non-linear relationship in the prediction of calories burn, which is physiologically correct.

This research emphasizes the need to assess the generalization abilities of models and leverage the use of physiological factors , especially in healthcare-ML related applications, which has been overlooked in previous research by merely relying on overall error rates.

7 Code & Data Availability Statement

The adopted dataset of all the research questions can be accessed at Calories Burnt Dataset | Kaggle.
This study's final code and datasets can be accessed here, CalFit | GitHub.

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 *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, pp. 1–4. Raipur, India. doi: 10.1109/ICPC2T53885.2022.9776710.

[2] Kadam, A., Shrivastava, A., Pawar, S.K., Patil, V.H., Michaelson, J. & Singh, A. (2023) Calories burned prediction using machine learning. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 1712–1717. Gautam Buddha Nagar, India. doi: 10.1109/IC3I59117.2023.10397623.

[3] Prasad, A., Asha, V., Vasumathi, M., Gupta, A., Nath, A. & Gehlot, A. (2025) Comparing gradient boosting and linear models for calorie prediction. In *2025 International Conference on Data Technology (ICDT)*, pp. 301–304. doi: 10.1109/ICDT63985.2025.10986394.

[4] Reddy, K.H.V. & Fernandez, T.F. (2024) Implementing calorie burnt prediction through XGBOOST algorithm compared over random forest regressor. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–4. Kamand, India. doi: 10.1109/ICCCNT61001.2024.10726282.

[5] Grammarly Inc. (2025) *Grammarly AI Writing Assistant* [software tool]. Retrieved from `https://www.grammarly.com`. Used during the final report drafting process for spell-checking and overleaf formatting purposes.

Appendix

A Learning Curve Tables

Table 8: Learning curve: LASSO (Polynomial)

Train Size	Train RMSE	Val RMSE	Gap (Val-Train)
960	2.745013	2.791172	0.046159
2194	2.774701	2.781012	0.006312
3428	2.783532	2.773602	-0.009930
4662	2.781620	2.769219	-0.012401
5897	2.767834	2.773334	0.005501
7131	2.765521	2.769275	0.003754
8365	2.759853	2.765528	0.005675
9600	2.762748	2.766212	0.003464

Table 9: Learning curve: Monotone GBDT

Train Size	Train RMSE	Val RMSE	Gap (Val–Train)
960	1.637566	4.130415	2.492850
2194	1.409126	2.819357	1.410231
3428	1.328837	2.386075	1.057238
4662	1.290506	2.069356	0.778850
5897	1.298409	1.916068	0.617659
7131	1.263901	1.833230	0.569330
8365	1.263862	1.771918	0.508057
9600	1.254355	1.682441	0.428085

Table 10: Learning curve: LASSO (Linear)

Train Size	Train RMSE	Val RMSE	Gap (Val–Train)
960	7.630672	7.638707	0.008035
2194	7.720428	7.603186	-0.117242
3428	7.794764	7.603944	-0.190819
4662	7.797658	7.596857	-0.200800
5897	7.770086	7.601146	-0.168939
7131	7.800877	7.600462	-0.200415
8365	7.790513	7.597767	-0.192747
9600	7.802242	7.599326	-0.202916