FairPlay Algorithm (FPA): Workout Status Analysis in Health Breath Classification

Jalen Joni, Ryan Moody, Maksym Grygorkiv, Phillip Olla, and Mina Maleki

Dept. of Electrical & Computer Engineering & Computer Science
University of Detroit Mercy
Detroit, MI, United States
(jonija, moodyrp, grygorma1, ollaph, malekimi)@udmercy.edu

***Abstract*—This project examines whether breath sensor data can identify whether a sample was collected before or after a workout. Using 242 aggregated breath tests from college athletes, the data was cleaned, normalized, and combined with demographic and survey information before applying machine learning and fairness analysis. The workflow included feature engineering, class balancing, and model evaluation to determine which attributes best support classification. The results show that gases such as acetone, CO2, and temperature provide meaningful separation between pre and post workout samples, and that balanced performance across demographic groups is achievable with proper preprocessing. These findings highlight the potential of breath based signals for monitoring athlete readiness and post exercise physiological changes.**

***Index Terms*—breath analysis, machine learning, classification, athlete monitoring, fairness**

I. Introduction

Breath analysis offers a noninvasive way to observe physiological changes that occur before and after exercise. Sensors capture gases such as acetone, CO2, hydrogen, methane, humidity, and temperature during a controlled exhalation, and these signals often shift with physical activity [1]. This makes breath data a promising tool for athlete monitoring, recovery assessment, and performance evaluation without the need for invasive tests. The dataset used in this project contains 242 breath samples from collegiate athletes combined with demographic information and wellness survey responses. Despite the richness of the sensor data, the dataset presents several challenges. The pre workout and post workout classes are imbalanced, demographic details are incomplete for some athletes, and hydration responses are heavily skewed toward recent fluid intake. These issues introduce noise and bias that require structured preprocessing, feature engineering, and class balancing techniques.

The goal of this project is to classify each breath sample as pre- or post-workout based on eight breath-gas metrics, demographics (gender, race, sport), and hydration timing. The classification algorithm outputs a binary prediction indicating whether the sample reflects pre- or post-exercise. The study also evaluates how consistently these models perform across demographic groups. The workflow includes data cleaning, normalization, demographic integration, engineered features, visualization, SMOTE-based class balancing, feature selection, and supervised learning models. Fairness is assessed by comparing model performance across demographic subgroups and examining whether balancing strategies improve consistency. The remainder of this paper is structured as follows: the framework section outlines the overall workflow; the methodology section describes data preparation, feature engineering, balancing methods, and model construction; the results and discussion section presents exploratory findings, classification outcomes, and fairness analysis; the conclusion summarizes key insights, limitations, and steps for future work.

II. Framework

The framework section provides an overview of the workflow used in this study. Figure 1 summarizes the complete process from raw data collection to model evaluation.

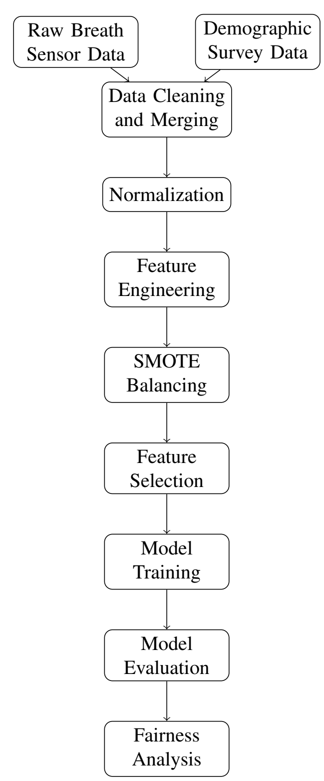

Fig. 1. Workflow for preparing and analyzing the breath data

III. Related Work

One relevant line of prior research examines how volatile organic compounds (VOCs) in exhaled breath respond to exercise. Bell et al. [3] evaluated the reliability of exhaled acetone and isoprene measurements following submaximal treadmill exercise in a controlled laboratory environment. Their study involved ten healthy active males with an average age of 22.7 years, a population similar in age and fitness level to the collegiate athletes analyzed in the present work. The authors collected breath samples before exercise, ten minutes after exercise, and twenty-four hours later, quantifying VOC concentrations using gas chromatography–mass spectrometry. Their findings showed that isoprene exhibited strong reliability across repeated exercise sessions and demonstrated a consistent post-exercise decrease, whereas acetone showed poor reliability, high variability, and no statistically significant change across time points. Although Bell et al. provide evidence that certain breath biomarkers respond predictably to exercise, their study is limited by a small, homogeneous sample and focuses only on two VOCs measured under tightly controlled conditions. In contrast, the present work analyzes a substantially larger and more diverse athletic population and incorporates eight gas-channel measurements collected through portable sensors in a real-world testing environment. Moreover, whereas prior work generally examines raw VOC concentrations, our study extends this line of research by evaluating engineered features such as the Acetone/CO_2 and Methane/Hydrogen ratios, which have not been explored in the exercise literature. This study further differs by applying supervised learning to classify pre- and post-workout states and by examining how performance varies across demographic groups, an area largely unaddressed in breath-analysis research. Together, these distinctions situate our work as a broader and more applied investigation of how multivariate breath data can characterize physiological changes surrounding exercise.

In addition to reliability studies, other research has examined how specific breath biomarkers reflect metabolic activity during exercise. Kim et al. [4] investigated whether breath acetone could estimate exercise-induced substrate utilization in young men running for one hour at 70% of VO_2max after a controlled low-carbohydrate diet. Breath acetone was measured alongside blood ketones and respiratory exchange ratio (RER), and the authors reported significant correlations between acetone and both β-hydroxybutyrate (r = 0.68) and RER (r = 0.67). These results support acetone as a metabolic indicator, showing that it rises with prolonged exercise and increased fat oxidation.While Kim et al. link acetone to metabolic changes, their work focuses on a single gas collected under controlled laboratory and dietary conditions. The present study differs by analyzing eight breath-gas measurements from a more diverse athletic population performing real-world workouts. Rather than examining acetone in isolation, this work incorporates engineered relational features such as the Acetone/CO_2 ratio and uses them for classifying workout state (pre versus post exercise) and evaluating subgroup consistency, representing a distinct application of breath analysis within exercise science.

Class imbalance is a well-documented challenge in biomedical classification problems, where minority classes often represent clinically important outcomes. Batista et al. [5] conducted an extensive evaluation of multiple imbalance-handling techniques across several medical datasets, comparing random oversampling, random undersampling, and the Synthetic Minority Oversampling Technique (SMOTE). Their results showed that SMOTE frequently improved minority-class recall without substantially degrading overall accuracy, particularly in algorithms sensitive to skewed class distributions. The authors also noted that different resampling strategies can alter the learned decision boundaries, emphasizing that imbalance correction techniques must be selected carefully based on the modeling objective. The present work mirrors this line of research through its comparison of a class-weighted Random Forest trained on the original dataset and a SMOTE-balanced version trained for exploratory purposes. While Batista et al. focus on medical diagnosis settings, the underlying issue is similar: physiological data often exhibit natural class imbalance. By analyzing both weighted and SMOTE-balanced models, the current study evaluates how imbalance-handling choices affect pre– and post-workout classification, ultimately opting for class weighting to preserve the natural distribution of breath samples in the final model.

Hydration status plays a critical role in exercise physiology and metabolic responses. In a comprehensive review, Périard et al. [6] describe how fluid loss during prolonged exercise under heat stress leads to reductions in plasma volume, elevated core temperatures, increased heart rate, and ultimately decreases in performance once body-mass losses approach 2% or more. Although this work does not directly measure breath-gas biomarkers, it underscores the mechanistic link between hydration and physiological shifts—shifts that are likely to alter exhaled gas compositions and their interpretive value. In the current study, hydration timing (Question 5) was included precisely because hydration-mediated physiological changes can modulate breath-gas signals and classification performance. Hence, our integration of hydration timing extends the exercise–breath-gas literature by addressing a contextual factor that prior breath-sensor studies have largely ignored.

Feature selection is a critical step in biomedical machine learning. Noroozi et al. [7] evaluated the impact of sixteen distinct feature-selection methods (filter, wrapper, evolutionary) on heart disease prediction, finding that while some methods improved accuracy by over 2%, others reduced classifier performance depending on the algorithm and number of selected features. Their results highlight that feature-selection strategies must be chosen carefully and tailored to the classifier and data context. In the present work, rather than aggressively reducing features, we computed mutual-information scores and inspected feature importance rankings, then retained the full feature set because empirical tests indicated no performance gain from removal.

IV. METHODOLOGY

A. Data Preparation

The dataset was provided as individual CSV files, each containing time-series breath measurements for a single test. Each file included eight gas-related attributes: Pressure, Acetone, Methane, Hydrogen, Contaminants, CO_2, Temperature, and Humidity. To create a single tuple for each sample, the mean value of each attribute was computed across the time series, resulting in one aggregated value per feature.

All aggregated samples were then combined into a unified dataset. Minor naming adjustments were made for clarity, including renaming `T_exhaust` to "Temperature" and relabeling the binary workout indicator as "Pre" and "Post". Redundant variables and fields no longer meaningful after aggregation were removed. Samples under the test subject IDs were excluded.

Samples containing incomplete demographic information were retained to preserve sample count, but demographic variables were omitted for those particular cases. No missing values were present in the gas measurements. Because the gas attributes varied in scale, min–max normalization was applied to all eight measurement features to create a consistent numerical range for analysis. The normalization was performed using the standard min–max formula:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

B. Feature Engineering

The feature set included normalized gas measurements, demographic variables, hydration timing (Question 5), and engineered attributes designed to capture relationships not expressed directly by individual gases. Three engineered features were constructed.

The first feature was the Acetone/CO_2 ratio,

$$\text{Ratio}_{A/CO_2} = \frac{\text{Acetone}}{\text{CO}_2},$$

which highlights instances where acetone is elevated relative to CO_2, a pattern that differed between pre and post workout states.

A second feature, the Methane/Hydrogen ratio,

$$\text{Ratio}_{M/H} = \frac{\text{Methane}}{\text{Hydrogen}},$$

captures the relative balance between methane and hydrogen, which often change together during physical activity.

The third engineered feature was the Combined Gas Index (CGI),

$$\text{CGI} = \frac{\text{Acetone} + \text{Methane} + \text{Hydrogen} + \text{CO}_2}{4},$$

which summarizes the overall gas environment in a single composite measure.

C. Handling Class Imbalance

The dataset contained more post workout samples than pre workout samples, creating a class imbalance that could bias models toward predicting the majority class. To address this without altering the data distribution, the final Random Forest classifier was trained using class weights that penalized errors on the minority class more heavily. This encouraged the model to recognize pre workout samples more reliably while preserving the natural proportions of the dataset.

An additional exploratory model was trained using Synthetic Minority Oversampling Technique (SMOTE) applied only to the training set. SMOTE generates new synthetic minority samples by interpolating between existing ones:

$$x_{\text{new}} = x + \lambda(x_{\text{nn}} - x), \quad \lambda \sim U(0,1).$$

This comparison allowed evaluation of how synthetic balancing affected recall and feature behavior. The primary results and conclusions are based on the class-weighted Random Forest trained on the original imbalanced dataset.

D. Feature Selection

To evaluate how strongly each attribute contributed to distinguishing pre workout and post workout states, mutual information scores were computed for all gas measurements, engineered features, hydration timing, and demographic indicators. Mutual information measures the reduction in uncertainty about the class label when a feature is observed:

$$MI(X,Y) = H(Y) - H(Y \mid X),$$

where $H(\cdot)$ denotes entropy. Larger values indicate stronger dependence between the feature and the target label.

The mutual information ranking showed that Humidity, the Acetone/CO_2 ratio, Race_Black Or African American, CO_2, Gender_female, and Race_Hispanic Or Latino were among the strongest contributors. Question 5 (time since last drink) also showed moderate influence, consistent with hydration-driven shifts in gas signals.

The individual scores produced by the model are summarized below:

Humidity	=	0.0923
Acetone/CO_2 ratio	=	0.0610
Race_Black Or African American	=	0.0550
CO_2	=	0.0305
Gender_female	=	0.0235
Race_Hispanic Or Latino	=	0.0189
Question 5 (hydration timing)	=	0.0184
Acetone	=	0.0127

These numeric values indicate how much each feature reduces uncertainty about whether a breath sample is pre workout or post workout. For example, Humidity has a score of 0.0923 while Acetone has a score of 0.0127. A larger value means that the feature contains more useful information for predicting the class label. In practical terms, a feature with a score around 0.09 contributes noticeably to classification, while a feature with a score near 0.01 contributes only a small amount. Features with scores of zero do not reduce uncertainty at all and therefore provide no measurable predictive value for the model. These values do not represent percentages or error rates; instead they indicate how strongly each feature is associated with the pre workout or post workout label relative to the others.

Several raw sensor channels showed negligible contribution, including Hydrogen, Methane, Pressure, and Contaminants, whose scores were zero. These attributes were retained because the Random Forest model can down weight uninformative variables automatically, and removing them did not increase accuracy or recall.

E. Machine Learning Models

Three supervised learning models were evaluated: decision trees, k nearest neighbors, and Random Forests. Decision trees tended to overfit the training data and showed sensitivity to the class imbalance. The k nearest neighbors classifier relied on Euclidean distance in the normalized feature space and performed inconsistently across runs, especially on minority class recall.

The Random Forest model offered the most stable performance due to its ensemble structure. A Random Forest constructs multiple decision trees, each trained on a bootstrap sample, and aggregates their predictions by majority vote. For a forest with T trees, the prediction is

$$\hat{y} = \text{mode}\left(f_1(x), f_2(x), \ldots, f_T(x)\right).$$

The final model used $T = 200$ trees with class weighting. The weighted loss increased the cost of misclassifying pre workout samples. For a class c, the effective weight applied during training was

$$w_c = \frac{N}{2 \cdot N_c},$$

where N is the total number of training samples and N_c is the number of samples belonging to class c. This weighting reduced bias toward the majority post workout class.

A secondary Random Forest was trained on a SMOTE balanced version of the training set. SMOTE synthesized new minority samples using

$$x_{\text{new}} = x + \lambda(x_{\text{nn}} - x), \qquad \lambda \sim U(0,1),$$

where x_{nn} is a nearest neighbor of x. This model allowed comparison between algorithmic reweighting and synthetic oversampling.

F. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1 score, and confusion matrices. These metrics characterize both overall correctness and behavior on the minority class.

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

The F1 score, the harmonic mean of precision and recall, is:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Macro-averaged precision, recall, and F1 were used so that each class contributed equally to the final scores. Confusion matrices were examined to understand error patterns, particularly false negatives on pre workout samples. All metrics were applied consistently to both the class-weighted model and the SMOTE-based comparison model.

V. Results and Discussion

A. Dataset

The final dataset contained 242 aggregated breath samples collected before and after exercise. Each sample included the mean values of seven gas sensor measurements (acetone, CO2, hydrogen, methane, humidity, pressure, and temperature) derived from the original time-based breath files. These were merged with demographic and survey fields to create a single integrated dataset suitable for supervised learning and analysis. The target variable consisted of two classes: pre workout and post workout. The distribution was uneven, with 141 post workout samples and 101 pre workout samples, indicating a moderate class imbalance. This distribution is shown in Table I.

TABLE I
Pre vs Post Workout Sample Distribution

Class	**Count**
Pre-Workout	101
Post-Workout	141
Total	242

Gender information included 110 male samples, 96 female samples, and 36 entries with no reported gender. This distribution is summarized in Table II. Sport representation was similarly uneven, with the largest groups coming from track and field/cross-country (68 samples) and soccer (52 samples), followed by lacrosse (30), basketball (28), softball (15), and fencing (11). Thirty-six samples did not include a reported sport.

TABLE II
Gender Distribution of Breath Samples

Gender	**Count**
Male	110
Female	96
No Data	36
Total	242

Race distribution showed that most athletes identified as White or Caucasian (129 samples) or Black or African American (49 samples). Smaller groups included two or more races (12 samples), Hispanic or Latino (10 samples), Asian (3 samples), Middle Eastern (2 samples), and American Indian or Alaska Native (1 sample). Thirty-six entries did not report race, limiting the reliability of race-based subgroup analysis. The full race distribution is provided in Table III.

TABLE III
RACE DISTRIBUTION OF ATHLETES

Race	Count
White or Caucasian	129
Black or African American	49
Two or More Races	12
Hispanic or Latino	10
Asian	3
Middle Eastern	2
American Indian or Alaska Native	1
No Data	36
Total	242

All gas sensor features were normalized to a 0–1 range, with mean values between approximately 0.17 and 0.56 across the attributes. Several features exhibited notable correlations, including strong positive relationships between hydrogen and contaminants, acetone and temperature, and temperature and humidity. These relationships reflect underlying physiological and environmental patterns captured by the breath sensors and provide context for later feature selection and modeling results.

Overall, the dataset is characterized by class imbalance, missing demographic entries, uneven representation across sports and race categories, and correlated sensor features. These properties motivate the use of class balancing techniques and fairness-aware evaluation during model training and assessment.

B. Classification Results

The main goal of the classifier is to separate pre workout and post workout samples while still paying attention to the minority pre class. Two Random Forest configurations were evaluated on the same held out test set. The first model was trained on the original imbalanced data with class weights. The second model was trained on a SMOTE balanced version of the training set. Table IV reports the overall metrics and the error distribution for both models.

TABLE IV
PERFORMANCE OF RANDOM FOREST MODELS ON THE TEST SET

Model	Acc.	Macro Prec.	Macro Rec.	Macro F1
RF (no SMOTE)	0.689	0.688	0.651	0.651
RF + SMOTE	0.656	0.641	0.629	0.630

	RF (no SMOTE)		RF + SMOTE	
	Pred. Pre	Pred. Post	Pred. Pre	Pred. Post
Actual Pre	11	14	12	13
Actual Post	5	31	8	28

The baseline Random Forest reaches higher accuracy and slightly better macro precision, recall, and F1 than the SMOTE based model. The lower block of Table IV shows that the baseline model correctly identifies 11 of 25 pre workout samples and 31 of 36 post workout samples, while the SMOTE model trades some post accuracy for a small gain in pre recall.

C. Feature Importance Analysis

Feature relevance was examined from two complementary perspectives. The first perspective, presented earlier in the Feature Selection subsection, uses mutual information to evaluate how strongly each feature is associated with the pre/post workout label when considered independently. Mutual information does not depend on any specific classifier and measures the reduction in uncertainty about the class when a feature is observed.

The second perspective is model based and relies on the Random Forest feature importances. These importances capture how frequently and how effectively a feature is used to split the data within the ensemble of decision trees. Whereas mutual information evaluates each feature in isolation, the Random Forest importances reflect multivariate interactions and compensations among correlated features.

For a Random Forest with M trees, the importance of feature j is computed as the average decrease in Gini impurity contributed by splits on that feature:

$$\mathrm{Imp}(j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in T_m(j)} \Delta I_{m,t},$$

where $T_m(j)$ is the set of nodes in tree m that split on feature j, and $\Delta I_{m,t}$ is the reduction in impurity produced at node t. The resulting scores are normalised to sum to one, so they represent relative—rather than absolute—importance.

The Random Forest feature importance scores (rounded to three decimals) are:

$$\begin{aligned}
&\text{Humidity} && = && 0.103\\
&\text{Methane_Hydrogen_Ratio} && = && 0.088\\
&\text{Methane} && = && 0.083\\
&\text{Hydrogen} && = && 0.079\\
&\text{Combined_Gas_Index} && = && 0.075\\
&\text{Acetone_CO}_2\text{_Ratio} && = && 0.073\\
&\text{Temperature} && = && 0.069\\
&\text{Pressure} && = && 0.067\\
&\text{CO}_2 && = && 0.065\\
&\text{Acetone} && = && 0.061\\
&\text{Contaminants} && = && 0.055\\
&\text{Question 4} && = && 0.028\\
&\text{Race_White Or Caucasian} && = && 0.023\\
&\text{Race_Unknown} && = && 0.020\\
&\text{Gender_female} && = && 0.016\\
&\text{Race_Black Or African American} && = && 0.014\\
&\text{Gender_male} && = && 0.013\\
&\text{Question 5} && = && 0.010\\
&\text{Race_Two Or More Races} && = && 0.008\\
&\text{Race_Hispanic Or Latino} && = && 0.006\\
&\text{Race_Asian} && = && 0.003\\
&\text{Race_Middle Eastern} && = && 0.002
\end{aligned}$$

These importance values do not represent physical quantities or probabilities. They indicate how much the classifier relied on each feature across its internal decision paths. A score near 0.10 means the feature was used frequently and produced meaningful impurity reductions, whereas a score near 0.01 indicates limited contribution to the model's predictive decisions.

The ranking shows that almost all of the model's decision-making power comes from gas sensor attributes such as Humidity, Methane_Hydrogen_Ratio, and Methane. Demographic variables have very small importances, which suggests that the classifier primarily leverages physiological signals and does not rely heavily on group membership when differentiating pre and post workout samples.

VI. Conclusion

This study examined whether breath sensor data can distinguish between pre workout and post workout physiological states using supervised learning and fairness-aware evaluation. Using 242 aggregated breath samples, the data preparation process produced a structured feature set through normalization, engineered gas ratios, demographic encoding, and composite gas measures. The dataset presented several challenges, including class imbalance, missing demographic entries, and strong correlations among gas attributes, all of which shaped the modeling and evaluation strategy.

Across the models evaluated, the class-weighted Random Forest demonstrated the most stable and consistent performance. It achieved the highest accuracy and the most balanced precision, recall, and F1 scores, outperforming the SMOTE-balanced model. Synthetic oversampling increased recall for the minority pre workout class only marginally while lowering performance on post workout samples, indicating that reweighting the loss was a more effective strategy than generating synthetic data. Feature relevance analysis further showed that physiological gas features—particularly Humidity, the Methane/Hydrogen ratio, Methane, Hydrogen, and the Acetone/CO_2 ratio—were the strongest predictors, while demographic variables contributed minimally, suggesting that the classifier relied primarily on breath-based physiological patterns rather than group-specific attributes.

The study is limited by the small dataset size, incomplete demographic and survey fields, and the variability introduced by portable breath sensors, which constrain generalizability. Future work should expand the dataset, improve demographic completeness, incorporate temporal modeling of breath dynamics, and explore additional engineered features and fairness interventions. Overall, the findings demonstrate that portable breath analysis captures meaningful exercise-related physiological changes and has potential as a noninvasive tool for monitoring workout state and athlete recovery.

References

[1] J. King, A. Kupferthaler, K. Unterkofler, H. Koc, S. Teschl, G. Teschl, W. Miekisch, J. Schubert, H. Hinterhuber, and A. Amann, "Isoprene and acetone concentration profiles during exercise on an ergometer," *Journal of Breath Research*, vol. 3, no. 2, pp. 027006, 2009. doi: 10.1088/1752-7155/3/2/027006.

[2] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, no. 5–6, pp. 375–381, 2003. doi: 10.1080/08839510390219264.

[3] L. R. Bell, M. A. Myers, J. T. Harvey, D. Hennessy, R. L. Worn, M. P. Wallen, G. Davis, and B. J. O'Brien, "Reliability of Exhaled Acetone and Isoprene in Healthy Active Adult Responses to Submaximal Treadmill Exercise: A Pilot Study," *Journal of Science in Sport and Exercise*, 2025. doi: 10.1007/s42978-025-00327-x.

[4] M. J. Kim, S. H. Hong, W. Cho, D. H. Park, E.-B. Lee, Y. Song, Y. S. Choe, J. H. Lee, Y. Jang, W. Lee, and J. Y. Jeon, "Breath acetone measurement-based prediction of exercise-induced energy and substrate expenditure," *Sensors*, vol. 20, no. 23, p. 6878, 2020. doi: 10.3390/s20236878.

[5] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing medical datasets," in *Proceedings of the 2004 Brazilian Symposium on Artificial Intelligence*, 2004, pp. 524–533.

[6] J. D. Périard, T. M. H. Eijsvogels, and H. A. M. Daanen, "Exercise under heat stress: thermoregulation, hydration, performance implications, and mitigation strategies," *Physiology Reviews*, vol. 101, no. 4, pp. 1873–1979, 2021. doi: 10.1152/physrev.00038.2020.

[7] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Scientific Reports*, vol. 13, article 22588, 2023. doi: 10.1038/s41598-023-49962-w.