Using Machine Learning Algorithms to Detect Heart Rate Abnormalities
Results from the EndoMondo's user dataset
Madhoolika Chodavarapu and Uday Singla

1 Dataset

For this study, the dataset we have chosen is EndoMondo's Fitness tracking data[1]. This dataset provides detailed information about various users and their exercise patterns. The metadata includes exercise type, exercise duration, heart rate measurements over time, derived movement speed, and other data collected via sensors. This caught our attention because of its relevance to a growing issue: the abundance of unverified fitness advice on social media and the internet. Claims of burning 1,000 calories in a week or gaining 5 kg of muscle in a month often promote exercises that may not align with an individual's physical needs. With the increasing accessibility of gyms, training centers, and other fitness resources, many people undertake activities without consulting professionals, which can lead to internal harm rather than health benefits. Reports suggest a rise in heart-related issues, such as heart malfunctions or heart attacks, due to high-intensity or cardio-focused workouts. Given this context, we aim to analyze this dataset to investigate whether fitness apparel or devices could be designed to monitor heart rates and detect abnormalities during exercise.

The dataset comprises 956 unique users and 67,113 rows of sequential data. Each row contains features such as user ID, gender, sport type, latitude, longitude, altitude, timestamp, heart rate, derived speed, distance, and more.

In our initial data exploration, we calculated the number of users participating in each sport to understand the potential for building similarity-based relationships between users within the same sport shown in the graph below (Figure 1)

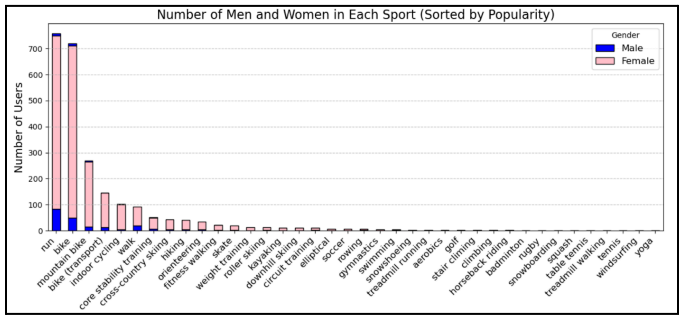

Figure 1: Distribution of users in sports

Additionally, we plotted the average heart rates of users against the sports they engage in to identify which sports might pose a higher risk for heart rate abnormalities (Figure 2).

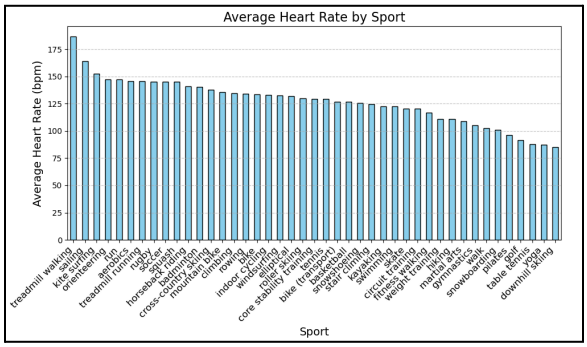

Figure 2: Average heart rates per sport

We further analyzed the data by grouping sports into five clusters based on heart rate patterns. This clustering helped us identify which groups exhibited the most heart rate variability—fluctuations in heart rate over time—and which activities generally elevated users' average heart rates (Figure 3).

[1] https://sites.google.com/view/fitrec-project/

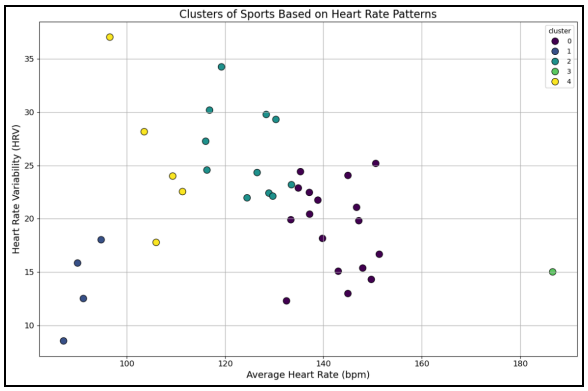

Figure 3: Clusters based on heart rate variability and bpm

We observed that the majority of sports with high average heart rates and significant variability are among the more popular activities in the dataset. This observation aligns well with our goal, as we hypothesize that heart rate abnormalities are often associated with rapid fluctuations, indicated by high variability. The popularity of these sports provides a substantial user base for monitoring and analysis.

However, we note that less than 15% of the data represent unpopular sports like "treadmill walking", which belong to the high-variability clusters. For such cases, where the user base is small but variability is high, it raises the possibility of outliers influencing our calculations. Although we have yet to determine the extent to which outliers impact variability, these cases suggest the need for further investigation. Our next step will involve mitigating the effects of outliers and conducting a detailed analysis of heart rate patterns for each sport to clean our data set for further use.

2 Predictive Task

Our task focuses on predicting potential heart rate abnormalities during exercise sessions. Through initial exploration, we observed that sports can be effectively clustered based on heart rate variability (fluctuations in bpm). This insight led us to include features influenced by the type of sport in our feature vector. Relevant features include **sport type**, **timestamp**, **heart rate**, **derived speed**, and **derived distance**. Additionally, we included **UserID** and **gender** as features since the dataset is unevenly distributed across genders, and physiological differences between genders significantly impact heart rate behavior.

The raw dataset did not originally include derived speed or distance. These features were calculated based on the methodology from the paper “Modeling Heart Rate and Activity Data for Personalized Fitness Recommendation”. **Derived distance** (in kilometers) was computed using the Haversine formula to calculate differences in latitude and longitude. **Derived speed** (in kilometers per hour) was obtained by combining derived distance and timestamps.

To explore heart rate abnormalities, we plan to compare sequential heart rate data of a querying user with that of other users sharing similar metadata (e.g., gender, sport, speed). This analysis will utilize three models: Naive baseline model **(NB)**, Logistic regression model **(LR)** and Support Vector Machine **(SVM)** model.

For the validation of **NB** and **LR** models, we will split the dataset into training and testing sets, form heart rate thresholds based on the mean heart rates of each exercise, and evaluate whether the model accurately identifies anomalies in the test set. Performance will be measured using precision and F1 scores. For hyperparameter tuning, we will perform grid search to get our optimum parameters. Thresholds for the **NB** model will be determined based on guidelines and information from medical research papers i.e., normal heart rates according to medicine.

For the validation of our SVM model, we will use a combination of normal train-test

validation and time-based validation. The dataset will be split chronologically, with earlier data assumed to represent normal heart behavior. Later data will be used for validation, accounting for periodic heart rate increases due to exercise. Errors will be detected by comparing the flagged user's heart behavior with that of other users performing the same exercise, who represent "expected" heart rate patterns.

3 Predictive Model

In our predictive analysis, we aimed to develop a model capable of detecting heart rate abnormalities during exercise sessions. Given the nature of our dataset—which includes sequential heart rate data alongside various user and activity features—we considered several modeling approaches before settling on our final models.

Since our dataset lacked the label of anomalies, we decided to create our own labels. We did this by grouping heart rate measurements by each user and the specific sport they indulge in. We made it context-specific because the mean heart rate may be different for different exercise-user pairs such as walking vs swimming. By grouping data by user and sport, we account for individual and activity-specific heart rate variations. We then calculated the mean heart rates and their standard deviation. We set thresholds where the normal ranges were set to be within 3 standard deviations from the average heart rate. If a heart rate was outside of these normal ranges, we marked it as an anomaly (label = 1); otherwise, it's normal (label = 0).

3.1 Naive Baseline Model

We began with a **Naive Baseline Model** that used global thresholds to identify anomalies. This model defined heart rate abnormalities based on deviations from the global mean heart rate across all users and activities. Specifically, we calculated the global mean (hr_mean_global) and global standard deviation (hr_std_global) of heart rates using the training data. An anomaly was defined as any heart rate measurement that deviated from the global mean by more than three standard deviations.

This simple threshold-based approach provided a quick way to label anomalies without considering individual differences or activity types. However, it lacked context sensitivity, leading to a high rate of false positives and negatives.

We chose the Naive Baseline Model as it provided a fundamental benchmark to evaluate the performance of more advanced models. By using global thresholds, it allowed us to quickly assess the basic effectiveness of anomaly detection without complex computations. This baseline was essential for comparing how much improvement was achieved with subsequent models. As we introduced logistic regression **(LR)** and Support Vector Machine **(SVM)**, we observed significant enhancements in prediction accuracy and a reduction in false positives and negatives. Ultimately, the Naive Baseline Model served as a crucial reference point, demonstrating the value of incorporating contextual information and more sophisticated algorithms in our anomaly detection approach.

3.2 Logistic Regression Model

To improve upon the naive baseline, we implemented a **Logistic Regression (LR) model** due to its simplicity, interpretability, and efficiency. Logistic regression is suitable for binary classification tasks and allows us to incorporate multiple features to better capture patterns associated with heart rate anomalies. Our features include: heart_rate, derived_speed (calculated using the Haversine formula to estimate exercise intensity), sport and gender encoded variables, heart rate difference from the

user-sport-specific mean, and the corresponding z-score.

To prevent data leakage, all statistical features were computed exclusively from the training data and then applied to the test set. Addressing class imbalance was achieved by setting class_weight='balanced', which prioritized the minority class of anomalies during training. We optimized the model through grid search over the regularization parameter and various solvers.

Despite these measures, the LR model exhibited high recall (83.78%) but low precision (6.41%), resulting in an F1 score of 11.91%. This indicates that while the model effectively identified most actual anomalies, it also produced a substantial number of false positives, which poses challenges for practical applications due to potential alarm fatigue.

3.3 Support Vector Machine Model

To address the limitations of the logistic regression model, we implemented a Support Vector Machine (SVM) model, leveraging its ability to handle non-linear relationships and perform effectively in high-dimensional spaces. The SVM utilized the same features as the logistic regression model, including heart rate, derived speed, encoded sport and gender variables, heart rate difference from the user-sport-specific mean, and the corresponding z-score. Consistent with the previous model, all variables were calculated using only the training data to ensure fair comparison.

We employed a radial basis function (RBF) kernel to capture complex patterns, set class_weight='balanced' to address class imbalance, and standardized all features with StandardScaler to ensure uniform feature contribution. Training the SVM on our large dataset was computationally intensive, which we managed by utilizing powerful computational resources, optimizing data structures, implementing batch processing, and appropriately setting convergence criteria.

Hyperparameter tuning was conducted through grid search, exploring various values for the regularization parameter C and the kernel coefficient gamma to achieve the optimal balance between complexity and generalization. The SVM model demonstrated superior performance, achieving a precision of 86.06%, recall of 99.97%, and an F1 score of 92.49%, significantly outperforming both the logistic regression and baseline models by effectively identifying anomalies while minimizing false positives and negatives.

3.4 Comparison of Models

In comparing the three models, each exhibited distinct strengths and weaknesses. The Naive Baseline Model was appreciated for its simple implementation and its role in providing a basic benchmark. However, it fell short by ignoring individual differences and activity types, resulting in high rates of false positives and negatives. The Logistic Regression Model offered improvements by incorporating multiple features and achieving higher recall compared to the baseline, while also being quick to implement. Despite these advantages, it struggled with low precision due to class imbalance and was limited in capturing non-linear patterns, making it vulnerable to data leakage if not meticulously managed. In contrast, the Support Vector Machine (SVM) Model excelled by capturing complex, non-linear relationships and achieving both high precision and recall, significantly enhancing overall performance. The main drawbacks of the SVM were its computational intensity and the substantial resources and time required for training. Overall, while the Naive Baseline provided a foundational benchmark, the Logistic Regression and SVM models demonstrated progressively enhanced capabilities, with the

SVM offering the best performance at the cost of increased computational demands.

3.5 Alternative Models and Unsuccessful Attempts

We considered implementing a Context-Aware Baseline Model that used user-sport-specific thresholds to define anomalies. However, we encountered data leakage because the model's thresholds overlapped with the labeling logic, leading to an unfair advantage. Additionally, the model did not offer significant benefits over the logistic regression model, prompting us to exclude it from our final analysis.

Other models we considered but did not implement included Random Forests and K-Nearest Neighbors (KNN). We decided against these due to concerns about overfitting and computational scalability with our large dataset.

4 Literature

4.1 Original usage

This dataset we used was originally extracted by Professor Julian McAuley and his team at UCSD in 2019 from EndoMondo's user activity logs. This data was employed in their research to develop *FitRec*, a Long Short-Term Memory (LSTM)-based recommender system designed to recommend fitness activities tailored to users. The researchers analyzed approximately 250,000 user activity records to model personalized activity dynamics by studying heart rate patterns and other features. Their work demonstrated the potential of deep learning models to predict and recommend fitness routines based on individual physiological and activity-based data, providing a foundation for more personalized and effective fitness guidance systems.

4.2 Usage of similar datasets

A similar paper to ours is “Healthcare and Anomaly detection: using machine learning to predict anomalies in heart rate data”[2] written in 2020. Although we have not taken inspiration from this paper, we found similarities in models and approaches taken which will be discussed ahead. This paper raises a similar question that we did: what is considered an anomaly? They classified anomalies as either rarely occurring data or non-common data. Both this research paper and our exploration considered anomalies of the latter kind.

The dataset they used for training and validation of their model was simulated: this meant that they could synthetically form labels and unlabeled data to see how well their models were performing. According to them, having control of distribution (between anomaly points and normal points) was advantageous in the training model. Their simulated data was also in the format of sequential heart rates, similar to the data we had.

To test their models, they had used MIT BIH[3] database that closely matched general heart rate patterns of those using heart rate tracking products - also formatted as sequential data. To decrease the rate of false alarms i.ie., False Positives, the paper explores multiple models and sets a few baselines - similar to what we did. Except, their baseline was extracted at a normal heart rate rate 60-100 bpm as stated by medical norm. In their approach, this made sense because the heart rate is not supposed to increase abnormally during normal activities. However, in our case, we needed to set a threshold that takes into account that the user is exercising, and also that different exercises cause different levels of heart rate increase. Hence we used a mean for each exercise and standard deviations

[2] https://link.springer.com/article/10.1007/s00146-020-00985-1
[3] https://ecg.mit.edu/time-series/

to set lower and upper bound heart rate thresholds for each exercise.

As for features, this paper had formed 3 new features out of the raw data: difference between current heart rate and previous heart rate, difference between current heart rate and the average of the past five heart rates (moving average) and k-means cluster number based on moving average. Based on these, the models they had used were: Local outlier factor (lof) - that assess how isolated a datapoint is with respect to its neighbours; Isolation Forests (IF) - that uses an isolation function that partitions data points until each of them are isolated. The number of steps taken to isolate a point determines its 'abnormality'; lastly, an SVM was used to optimally draw a boundary between different classes of data to decrease misclassification error. In our feature expansion, we also formed a feature that represents the mean of heart rate, except it was across users during the same time stamp - that is we take the average of heart rates at every time stamp across all users. This was more relevant to our study since our user's heart rate increases as the time into their exercise increases. As for validation, similar to us, they had also used k-fold cross validation (k=5) to test out different parameters to evaluate performance. For evaluation, their ground truth was based on their rule of thumb 6-100 bpm as mentioned earlier.

4.3 Main conclusions in the literature

Both our model and the literature model aim to classify heart rates as either normal or abnormal. A key difference lies in the nature of the data: their model uses static heart rate data from monitors, while ours uses dynamic heart rate data captured during exercise sessions.

The study from the literature evaluated five algorithms for heart rate anomaly detection, analyzing performance differences based on the proportion of anomalies in the training dataset (0.5% vs. 2.5%). On simulated data, SVM excelled in scenarios with fewer anomalies, whereas Random Forests performed best with higher anomaly rates. The LOF algorithm emerged as particularly promising for real-world applications due to its conservative classification approach, effectively detecting anomalies across both low and high heart rate ranges.

A significant takeaway from the research is the efficacy of training models on simulated data when real labeled data is unavailable. This rule-based approach provides a valuable framework for tuning anomaly detection systems in their early stages, adapting them to real-world scenarios.

5. Results and Conclusions

Our analysis demonstrated significant improvements in anomaly detection performance as we progressed from the Naive Baseline Model to the Logistic Regression (LR) and Support Vector Machine (SVM) models. The Naive Baseline Model served as a fundamental benchmark, achieving a precision of 10%, a recall of 40%, and an F1 score of 16%. These metrics highlighted the baseline's limitations, with low precision and recall stemming from its inability to account for individual differences and activity-specific heart rate variations. Consequently, the baseline model resulted in a high rate of false positives and negatives, underscoring the necessity for more sophisticated approaches.

In contrast, the Logistic Regression Model improved recall to 83.78%, effectively identifying a majority of actual anomalies. However, this model was plagued by a low precision of 6.41%, indicating a substantial number of false positives. This imbalance arose from the class weighting and the model's linear

nature, which struggled to capture the complex, non-linear patterns inherent in the heart rate data. The low precision poses practical challenges, such as alarm fatigue, where users may become desensitized to frequent false alerts.

The Support Vector Machine (SVM) Model markedly outperformed both the baseline and logistic regression models, achieving a precision of 86.06% and a recall of 99.97%, culminating in an F1 score of 92.49%. This superior performance can be attributed to the SVM's ability to handle non-linear relationships through the radial basis function (RBF) kernel and its effectiveness in high-dimensional feature spaces. By standardizing features and carefully tuning hyperparameters, the SVM successfully minimized false positives and negatives, providing a robust solution for anomaly detection in heart rate data.

Feature Representation Analysis revealed that incorporating derived features such as speed and heart rate difference (hr_diff) significantly enhanced model performance. The z-score, representing standardized heart rate deviations, was particularly effective in capturing anomalies relative to user-specific and activity-specific baselines. Encoding categorical variables like sport and gender also contributed to the models' ability to account for physiological and activity-induced heart rate variations.

Interpretation of Model Parameters highlighted the importance of feature scaling and the choice of kernel in the SVM. The RBF kernel enabled the SVM to create flexible decision boundaries, accommodating the intricate patterns in the data that linear models like logistic regression could not. Additionally, the balanced class weighting in both LR and SVM models ensured that the minority class of anomalies was adequately represented during training, thereby improving recall and precision.

In comparing models, the progression from the Naive Baseline to Logistic Regression and ultimately to SVM demonstrated the tangible benefits of incorporating contextual information and leveraging more sophisticated algorithms. While the baseline provided essential insights, the SVM delivered robust and reliable anomaly detection, making it the most effective model for our task. Despite its high recall, the Logistic Regression model was hindered by low precision, limiting its practical applicability.

Significance of the Results lies in the demonstration that advanced machine learning techniques, when appropriately configured and combined with thoughtful feature engineering, can substantially enhance the accuracy and reliability of heart rate anomaly detection systems. The superior performance of the SVM model underscores the importance of handling non-linear relationships and class imbalances in complex datasets.

In conclusion, the progression from a simple baseline to more sophisticated models underscored the necessity of incorporating contextual and derived features in anomaly detection tasks. While the Naive Baseline provided essential insights, the Logistic Regression model offered improvements in recall, and the SVM achieved outstanding overall performance by effectively balancing precision and recall. These results affirm that advanced machine learning techniques, when appropriately configured, can significantly enhance the accuracy and reliability of heart rate anomaly detection systems. Future work may explore ensemble methods or deep learning approaches to further refine detection capabilities and address any remaining computational challenges, thereby contributing to the development of more effective heart health monitoring systems in fitness applications.