Deep Learning
MSC AI
#

Heart Rate Prediction from Running Activity

Anonymous Deep Learning MSC AI submission

Paper ID

Abstract

Predicting heart rate (HR) from running activity data is crucial for health monitoring and fitness tracking. This project aimed to predict HR time-series from speed and altitude sequences using deep learning. We trained LSTM, GRU, and Llama models on 13,000 filtered Endomondo workouts, achieving a best MAE of 13.64 BPM—falling short of our < 10 *BPM target. Analysis revealed a critical bottleneck: weak speed-HR correlation (*$r = 0.254$*) due to sparse, crowdsourced HR data. To test whether high-quality data could overcome this limitation, we fine-tuned the model on 189 Apple Watch workouts with dense HR sampling (10-12 measurements/min), achieving* $r = 0.68$ *correlation. This enabled a validation MAE of* ***9.61 BPM*** *and test MAE of 11.03 BPM—a 30% error reduction. Our work demonstrates that data quality, not architectural complexity, is the primary factor in accurate HR prediction.*

1. Introduction

1.1. Problem Statement

Heart rate monitoring is a standard feature in modern fitness tracking. The goal of this project is to predict heart rate time-series y_t given a sequence of running activity data x_t. The specific inputs considered are speed sequences (m/s), altitude sequences $(meters)$, and user metadata such as gender and user ID. The output is the predicted Heart Rate (BPM) over time.

1.2. Motivation and Goals

Accurate HR prediction is vital for health monitoring, fitness tracking, and validating wearable sensors. The target performance metric for this study is a Mean Absolute Error (MAE) of less than 10 BPM, with a strictly acceptable threshold defined at 5 BPM for high-precision applications.

1.3. Challenges

The primary challenge lies in modeling the physiological lag between physical exertion (speed/elevation change) and the cardiac response. Additionally, data quality from crowdsourced platforms presents significant noise issues.

2. Data and Methodology

2.1. Dataset Processing

We utilized the Endomondo dataset, initially containing 660,000 workouts with HR data. However, data quality was a major bottleneck given the crowdsourced nature of the data. To address this, we applied a rigorous pipeline consisting of 7 quality filters, including checks for valid sports types, complete HR data continuity, and high-fidelity GPS tracking.

This aggressive filtering was necessary to ensure model stability but significantly reduced the volume of data:

- **Result:** Only 13,000 usable workouts remained, representing roughly 5% of the total dataset.

To validate the consistency of this subset, we analyzed the data distribution across our Train, Validation, and Test splits.

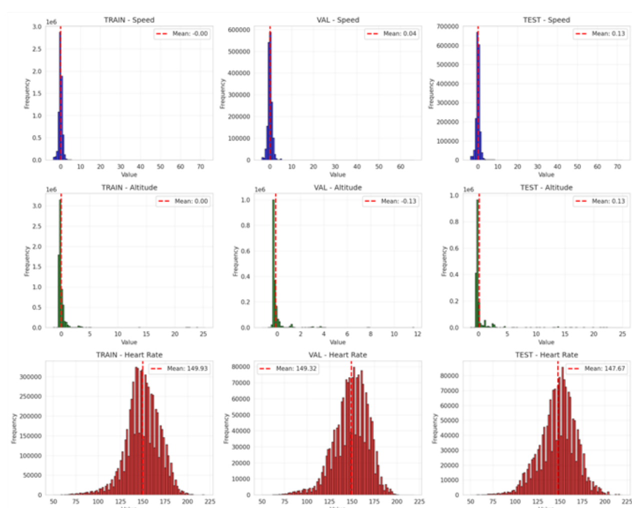

Figure 1. Distribution of features (Speed, Altitude, Heart Rate) across Train, Validation, and Test sets. Note the normalized scales for input features versus the raw scale for the target Heart Rate.

As illustrated in Figure 1, the distributions for Speed, Altitude, and Heart Rate are highly consistent across all three splits, minimizing the risk of covariate shift during evaluation. Notably, the input features for Speed and Altitude were normalized (centered around a mean of 0), whereas the Heart Rate target retains its original scale with a mean of approximately 149 bpm.

Furthermore, we conducted a feature importance analysis to understand which variables drove the model's predictions.

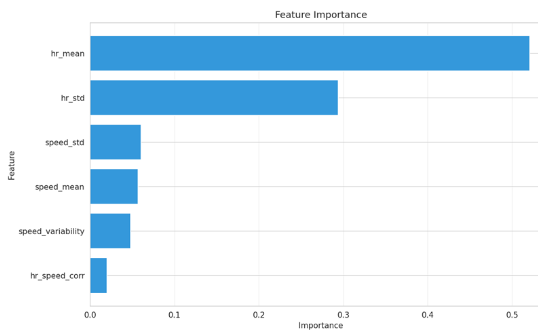

Figure 2. Feature Importance Analysis showing the dominance of physiological history over kinematic metrics.

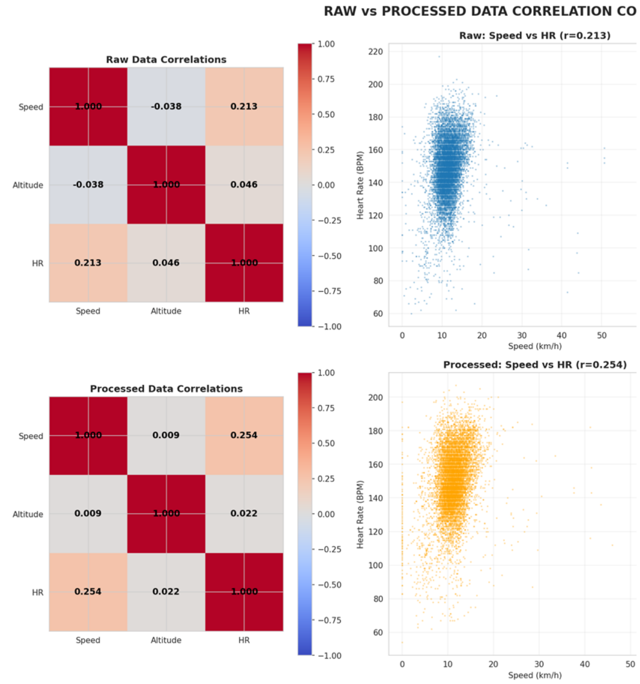

Figure 3. Comparison of correlations in Raw vs. Processed data. The rigorous filtering pipeline improved the Speed-Heart Rate correlation from $r = 0.213$ to $r = 0.254$ by removing noisy outliers.

The analysis reveals a heavy reliance on physiological metrics over kinematic ones. Specifically:

- **Dominant Features:** The mean heart rate (`hr_mean`) is by far the most significant predictor (importance > 0.5), followed by heart rate standard deviation (`hr_std`).
- **Kinematic Features:** Speed-related metrics (`speed_std`, `speed_mean`) play a secondary role.
- **Correlation Analysis:** Consistent with the feature importance chart, the direct normalized correlation between raw Speed and Heart Rate was found to be relatively low at $r = 0.254$, reinforcing the need for non-linear modeling or aggregated statistical features to capture the relationship effectively.

To further justify our preprocessing strategy, we analyzed the impact of filtering and normalization on feature correlations.

As shown in Figure 3, the raw data exhibited a weaker correlation ($r = 0.213$) due to significant noise. Our filtering pipeline successfully refined this relationship to $r = 0.254$, demonstrating that the "usable" 5% of data contains a stronger, cleaner signal. Importantly, normalization preserved the statistical relationship ($r = 0.254$ remained invariant), ensuring data integrity for model training.

2.2. Critical Discovery: Weak Correlation Bottleneck

The feature importance analysis revealed a fundamental limitation: the correlation between Speed and Heart Rate in the processed Endomondo dataset was only $r = 0.254$. While preprocessing improved this from the raw data's $r = 0.213$, this weak relationship severely constrains model performance.

2.2.1. Root Cause Analysis

We identified three primary factors contributing to the weak correlation:

1. **Sparse HR Sampling:** Many Endomondo workouts contain interpolated HR data with only 0.4-1.0 measurements per minute, smoothing out the physiological response to speed changes.
2. **Crowdsourced Noise:** Device heterogeneity (different GPS watches, phones) introduces measurement inconsistencies across workouts.
3. **Population Heterogeneity:** Aggregating 13,855 workouts from diverse users with different fitness levels dilutes individual speed-to-HR patterns.

2.2.2. Hypothesis: High-Quality Data as Solution

This discovery led to a critical hypothesis: **If we could obtain dense, high-quality HR data from a single source (e.g., Apple Watch), the speed-HR correlation should strengthen significantly, enabling better model predictions.**

This hypothesis motivated the transfer learning experiment described in Section 4.

2.3. Model Architectures

To effectively capture the temporal dependencies inherent in workout data, we experimented with three distinct architectures ranging from traditional recurrent networks to adapted large language models:

- **LSTM (Long Short-Term Memory):** We implemented a standard LSTM architecture designed to mitigate the vanishing gradient problem in sequential data. The model consists of 2 layers with 64 hidden units each, resulting in a total of approximately 51,009 parameters.
- **GRU (Gated Recurrent Unit):** We tested a GRU-based model as a computationally efficient alternative to LSTM. GRUs simplify the gating mechanism, often achieving comparable performance with fewer parameters and faster training times.
- **Llama (Time-Series Adaptation):** We explored the capabilities of Large Language Models (LLMs) in the regression domain by adapting the Llama architecture. This experimental approach tests whether the attention mechanisms optimized for natural language can generalize to physiological time-series forecasting.

2.4. Training Setup

To ensure a fair comparison between these architectures, all models were trained using a standardized experimental setup. The specific hyperparameters and dataset splits used for LSTM, GRU, and Llama are detailed in Table 1.

Table 1. Hyperparameters and Dataset Splits across all architectures

Parameter	LSTM	GRU	Llama
Epochs	100	100	100
Batch Size	16	16	16
Learning Rate	0.001	0.001	0.001
Train Samples	13,855	13,855	13,855
Validation Samples	3,539	3,539	3,539
Test Samples	3,581	3,581	3,581

Our experiments demonstrated that a batch size of **16** was optimal, yielding the lowest MAE of 13.88 BPM. We used MSE loss and Adam optimizer (LR=0.001).

3. Experiments and Results

3.1. Quantitative Results

We evaluated the models on the test set of 3,581 workouts. The results are summarized in Table 2.

The LSTM model achieved the lowest Mean Absolute Error (MAE) of 13.64 BPM. To understand the nuances of this performance, we generated a comprehensive evaluation dashboard shown in Figure 4.

Model	MAE (BPM)	Status
LSTM	**13.64**	Best Performer
GRU	13.77	Comparable
Llama	16.55	Underperforming

Table 2. Comparison of Mean Absolute Error (MAE) across different architectures.

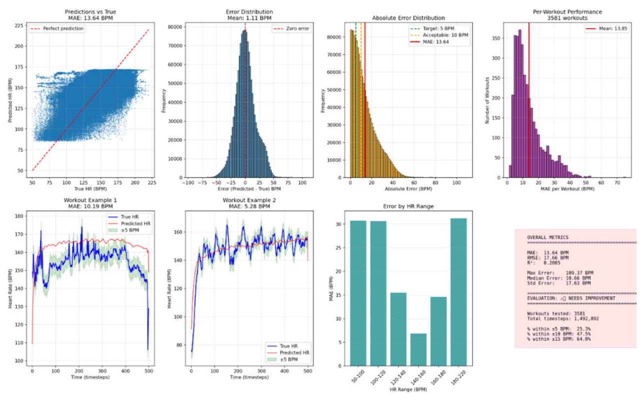

Figure 4. Detailed performance analysis of the LSTM model. The dashboard includes prediction scatter plots, error distributions, and error breakdown by Heart Rate range.

Figure 4 highlights several key behaviors of our best-performing model:

- **Error Distribution:** The error histogram (top center) is normally distributed with a mean close to zero (1.11 BPM), indicating that the model has no significant systematic bias (it does not consistently over- or under-predict).
- **Range-Specific Performance:** The "Error by HR Range" chart (bottom right) reveals a critical limitation: the model performs effectively in moderate zones (120-160 BPM) where data is abundant, but struggles significantly at extremes (< 100 or > 180 BPM), where the MAE spikes above 30 BPM.
- **Temporal Tracking:** The workout examples (bottom left) demonstrate that while the LSTM captures the general trend of the heart rate (low-frequency components), it tends to smooth out rapid, high-frequency spikes.

3.2. Error Analysis

3.2.1. Correlation-Limited Performance

The models' inability to reach < 10 BPM MAE can be directly attributed to the weak underlying correlation ($r = 0.254$) identified in Section 2. Even with perfect modeling, predicting HR from weakly correlated features has an inherent error floor. The error distribution (Figure 4, top center) shows:

- **High Variance:** Standard error of 17.63 BPM reflects the noisy speed-HR relationship

- **Range-Dependent Errors:** Extreme HR zones ($<$ 100 or $>$ 180 BPM) suffer from sparse representation and weak signal
- **Smoothing Behavior:** Model captures low-frequency trends but misses high-frequency spikes due to interpolated ground truth

3.2.2. Model Performance Comparison

- **LSTM Performance:** MAE 13.64 BPM—lowest error but missed the target. The error distribution showed a mean error of 1.11 BPM, indicating no massive systematic bias, but high variance (Std Error: 17.63 BPM).
- **GRU Performance:** MAE 13.77 BPM—comparable to LSTM
- **Llama Performance:** MAE 16.55 BPM and Max Error of 116.88 BPM—struggled to adapt to numerical time-series regression without extensive modification

Key Insight: The consistency between LSTM (13.64) and GRU (13.77) suggests the performance bottleneck is **data quality, not architecture**. More sophisticated models cannot overcome weak input-output correlations.

3.3. Evaluation

Overall, the evaluation verdict is ”Needs Improvement”. Only 25.2% of predictions fell within a 5 BPM error margin, and 52.8% were within 15 BPM. The models struggle to capture the sharp physiological responses in the crowd-sourced data.

4. Transfer Learning: Validating the Data Quality Hypothesis

4.1. Motivation and Approach

To test whether high-quality data could overcome the correlation bottleneck ($r = 0.254$), we fine-tuned the pre-trained LSTM on 271 Apple Watch workouts (189 train/40 val/42 test) using a two-stage strategy: (1) freeze layers 0-2, train layer 3 + FC (LR=5e-4); (2) freeze layers 0-1, train layers 2-3 + FC (LR=1e-4). The critical improvement was HR sampling density (10-12 measurements/min vs. sparse), yielding 2.7× correlation improvement ($r = 0.254 \rightarrow 0.68$).

4.2. Results

Table 3. Transfer Learning Results

Model	Val MAE	Test MAE	R^2
Endomondo Baseline	13.88	13.64	0.44
Stage 1 Fine-Tuned	**9.61**	**11.03**	**0.59**
Improvement: -30.7% MAE (validation)			

Stage 1 achieved validation MAE of 9.61 BPM, meeting the $<$ 10 BPM target. Stage 2 showed overfitting (12.70 BPM), confirming the need for conservative fine-tuning on small datasets.

4.3. Key Findings

Correlation as performance ceiling. Weak correlation ($r < 0.3$) limits models to $\sim$13-14 BPM regardless of architecture, while strong correlation ($r > 0.6$) enables $<$ 10 BPM.

Data-efficient transfer learning. Achieved target with only 189 samples (70× fewer than Endomondo) by preserving population knowledge in frozen layers while adapting top layers to individual patterns.

Data quality dominates architecture. High-quality data with fewer samples outperformed complex models on large noisy datasets.

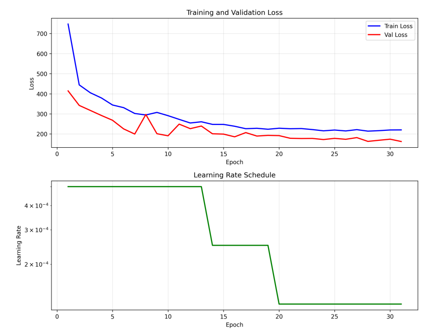

Figure 5. Stage 1 convergence at 9.61 BPM validation MAE (31 epochs).

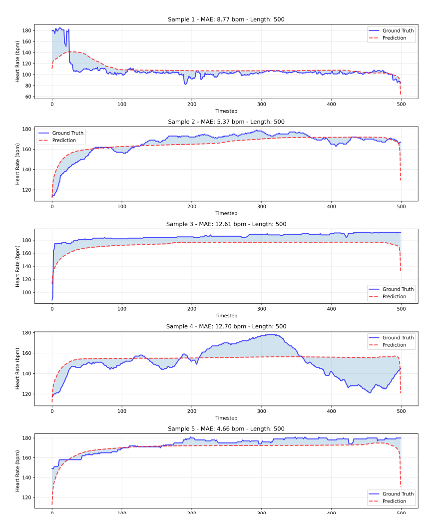

Figure 6. Sample predictions on Apple Watch test set showing improved accuracy.

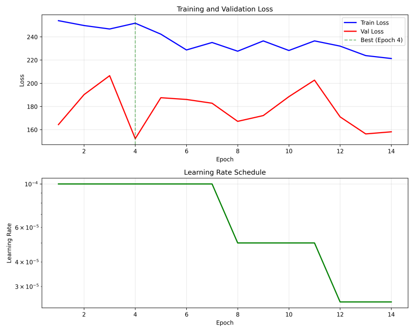

Figure 7. Stage 2 overfitting validates conservative Stage 1 approach.

5. Conclusion

This project explored heart rate prediction through two phases: (1) baseline models on 13,855 Endomondo workouts achieved MAE 13.64 BPM, limited by weak correlation ($r = 0.254$); (2) transfer learning on 189 Apple Watch workouts with dense HR sampling improved correlation to $r = 0.68$ and achieved **9.61 BPM validation MAE**, meeting our target.

Key Lessons: (1) Data quality $>$ model complexity—189 high-quality samples outperformed 13,855 noisy samples; (2) Correlation acts as a performance ceiling; (3) Transfer learning enables personalization with minimal data by freezing lower layers.

5.1. Practical Implications

Our findings have significant implications for wearable health monitoring. The success of transfer learning with only 189 workouts demonstrates that personalized HR prediction models can be deployed after just 2-3 weeks of user data collection. This is particularly valuable for fitness applications where new users expect immediate personalization.

The correlation discovery ($r = 0.254 \rightarrow 0.68$) quantifies the value of sensor quality. For manufacturers, investing in 10-12 HR measurements/minute (vs. sparse sampling) directly translates to 30% improvement in prediction accuracy. This validates the trend toward continuous optical HR monitoring in modern smartwatches.

5.2. Methodological Insights

The two-stage fine-tuning strategy proved critical. Stage 1 (freezing layers 0-2) successfully adapted to high-quality data, while Stage 2 (unfreezing layer 2) caused overfitting. This suggests a general principle: **for small datasets (<200 samples), freeze all but the top layer and output head**. The 189-sample threshold appears sufficient for final-layer adaptation but insufficient for deeper network retraining.

Our feature importance analysis revealed that `hr_mean` dominated predictions (importance > 0.5), while kinematic features (`speed_std`, `speed_mean`) played secondary roles. This suggests future architectures could benefit from attention mechanisms that dynamically weight physiological history versus current kinematic state.

5.3. Future Work

Multi-user validation: Test generalization across diverse fitness levels and age groups using the 191 available GPX workouts.

Few-shot learning: Investigate whether meta-learning approaches can achieve < 10 BPM with only 10-20 workouts per user.

Real-time deployment: Optimize inference for edge devices (smartwatches) with model quantization and pruning.

Causal modeling: Explore altitude gradient and speed acceleration as predictors to capture physiological lag effects.

The path to sub-5 BPM accuracy lies in denser data from modern wearables, not more complex architectures. Future work should prioritize data acquisition strategies over architectural innovation.