ISSN 2007-9737

An Advanced LSTM Model to Enhance Running Economy

José Luis Pasarin, Ulises Orozco-Rosas*, Kenia Picos

CETYS Universidad,
Mexico

joseluis.pasarin@cetys.edu.mx, {ulises.orozco, kenia.picos}@cetys.mx

Abstract. Running Economy (RE) is an important physiological measure for endurance athletes, particularly distance runners. It is defined as the energy demand for a specific velocity of submaximal running and depends on biomechanical, metabolic, cardiorespiratory, and neuromuscular factors. In this work, we introduce a novel approach for predicting running economy in amateur runners through the analysis and comparison of three neural network architectures: a proposed Long Short-Term Memory (LSTM) model, an alternative LSTM model, and a Recurrent Neural Network (RNN). Each model incorporates physiological and biomechanical metrics, including pace, heart rate, power, cadence, ground contact time, and stride length, collected with a wearable device. By analyzing these temporal sequences, the models estimate running efficiency using initial test data as a baseline. Their performance was evaluated using statistical metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The best results obtained from the proposed LSTM Model were a test MSE of 0.036995, a test MAE of 0.102474, and a test R^2 of 0.986986, demonstrating superior predictive capacity compared to the other architectures. These results highlight the potential of integrating artificial intelligence with wearable technology to provide accessible and personalized training tools for coaches and amateur athletes.

Keywords. Running economy, deep learning, LSTM network, time series, regression.

1 Introduction

Running economy is a key determinant of endurance performance and has been widely recognized as one of the most important physiological indicators for distance runners. Traditionally, its assessment has relied on controlled laboratory environments using indirect calorimetry and metabolic testing to quantify oxygen consumption. However, recent advances in bioenergetic modeling and data-driven estimation methods have begun expanding the possibilities for evaluating metabolic demand outside the laboratory. Emerging approaches—including novel analytical models of energy contribution during running [7] and deep learning–based estimators of metabolic rate derived from wearable biosignals [3] highlight a growing shift toward more flexible, non-invasive measurement techniques that complement or reduce the need for traditional metabolic equipment. However, these methods are costly, time-consuming, and often inaccessible to athletes and coaches outside laboratory environments, particularly given the need for specialized protocols and equipment highlighted in recent work on metabolic assessment and endurance performance [15, 17].

Recent advances in wearable technology and artificial intelligence have opened new possibilities for estimating key physiological parameters outside laboratory settings, with emerging machine learning approaches demonstrating the ability to predict aerobic capacity and performance indicators using non-exercise or field-based data [30]. Devices capable of recording continuous data—such as heart rate, power output, cadence, and stride metrics—enable the collection of large volumes of biomechanical and physiological information in real-world conditions. Deep learning approaches, particularly RNNs and LSTM architectures, have demonstrated strong capabilities in modeling sequential physiological signals, as evidenced by their successful application in energy-expenditure

estimation and activity-related metabolic prediction in prior studies [19, 25], making them well-suited for capturing the temporal dynamics underlying running performance. Despite these advancements, few studies have specifically examined the direct estimation of running economy using wearable sensor data in combination with temporal deep learning approaches, despite the growing body of research demonstrating the capabilities of machine learning and smart devices for health and activity monitoring [21, 22].

Furthermore, recent work in related domains highlights the expanding role of machine learning and sensor-driven systems in capturing, modeling, and interpreting complex human-generated data streams. Studies leveraging IoT-enabled architectures and time-series prediction models have demonstrated the feasibility of extracting meaningful behavioral and physiological patterns from continuous, device-based measurements, underscoring the value of sequential modeling techniques for non-laboratory monitoring [1].

Complementary research employing hybrid neural-network frameworks for emotion recognition from physiological signals further illustrates the potential of advanced learning algorithms to decode subtle, temporally evolving features in human biosignals [9]. Together, these developments reinforce the promise of applying deep learning to wearable-derived running metrics and support the need for specialized models capable of estimating running economy in real-world environments.

This research aims to address this gap by developing and evaluating data-driven models capable of estimating running economy in amateur runners using time-series data collected from wearable devices. Specifically, three neural network architectures are analyzed and compared: a proposed LSTM model, an alternative LSTM model, and an RNN. Each model processes temporal sequences of physiological and biomechanical variables to predict running economy values, using baseline test data for calibration. The comparison between these models provides insight into the influence of different architectural choices on prediction accuracy and generalization performance.

In summary, the main contributions of this work are:

— The proposed LSTM model incorporates physiological and biomechanical metrics, including pace, heart rate, power, cadence, ground contact time, and stride length, which are not considered in traditional mathematical formulations of running economy.

— While running economy is typically measured through laboratory testing, this work proposes a novel approach to estimate it using wearable sensor data and deep learning models, thus eliminating the need for laboratory equipment.

— The implementation and comparison of multiple recurrent neural network architectures, particularly the LSTM-based models, demonstrate their effectiveness in learning temporal patterns and accurately estimating RE from real-world running data.

The organization of this work is outlined as follows: Section 2 establishes the conceptual foundation by defining RE (running economy), discussing its physiological and biomechanical determinants, describing approaches for its estimation using both theoretical methods and wearable sensors, and introducing the deep learning principles used in this study, along with the evaluation methods applied to assess model performance. Building on this foundation, Section 3 describes the dataset preparation, the design of the deep neural network architectures, and the procedures used to train and validate the predictive models. It outlines the rationale behind each modeling approach and the criteria used to measure their effectiveness. Section 4 presents the experimental setup and examines how the models behaved during training and prediction. It evaluates their accuracy, analyzes their ability to represent temporal variations in RE, and offers a broader interpretation of their performance and limitations within real-world running contexts. Finally, Section 5 summarizes the key outcomes of the research and identifies future directions for advancing model accuracy and expanding the practical applications of data-driven RE assessment.

2 Theoretical Framework

Running has become one of the most prominent recreational and competitive sports activities worldwide, driven by growing interest in physical health, stress management, and performance optimization. As participation increases, scientific and academic research has focused on understanding the physiological and biomechanical determinants of running performance, particularly the efficiency with which individuals sustain prolonged effort.

Within this context, running economy emerges as a central indicator for evaluating and improving endurance performance.

2.1 Running Economy

Running economy has emerged as a central construct in endurance physiology due to its strong predictive value for performance across a wide range of running events. Consequently, its study has become essential for researchers and practitioners seeking to optimize training interventions, assess biomechanical adaptations, and develop performance-enhancing strategies. The following definition contextualizes this concept within the scope of endurance performance research.

Definition: Running economy refers to the energy cost required to sustain a given running speed, commonly assessed through oxygen consumption (VO_2) during submaximal exercise.

Recent analyses have emphasized the importance of accurately quantifying this metabolic demand to better evaluate endurance performance [5, 14]. A lower energy cost for a given running velocity reflects better running efficiency, which is strongly associated with enhanced performance in middle- and long-distance events [23, 28].

From a physiological perspective, RE reflects metabolic efficiency, cardiopulmonary capacity, and the ability to transport and utilize oxygen efficiently [5]. Biomechanically, RE depends on neuromuscular coordination, musculoskeletal function, and the effectiveness with which mechanical energy is converted into forward propulsion. Key intrinsic factors influencing RE include stride length, cadence, running technique, and elasticity of the musculotendinous system [4, 28].

RE serves as a critical variable to monitor athletic performance, guide training strategies, and support injury prevention. Advances in wearable technology and data-driven modeling have expanded the ability to estimate RE in real-world environments, enabling individualized feedback for continuous performance improvement.

Determinants of Running Economy: RE results from a multifactorial interaction between physiological and biomechanical components.

Biomechanical variables such as cadence, stride length, ground contact time, running power, and vertical oscillation show direct influence on running efficiency [5, 28]. As summarized in Table 1, higher cadence, reduced vertical oscillation, and shorter ground contact time are associated with lower metabolic cost and improved RE.

Wearable sensors now enable the accurate measurement of these parameters in natural training environments, allowing for large-scale collection of gait and physiological data [10, 26].

Despite the high validity of metrics such as cadence and stride duration, specific parameters like braking force and vertical displacement may exhibit variability and require specialized analysis to ensure accuracy [10].

Understanding the determinants of RE is essential for designing personalized training programs, optimizing mechanical efficiency, and reducing injury risk in both recreational and competitive runners. Recent advancements further highlight that these determinants are closely linked to metabolic cost prediction, as demonstrated by emerging sensor-based modeling approaches [29].

2.2 Running Economy Estimation

While running economy can be empirically quantified through laboratory-based assessments, its evaluation often benefits from complementary theoretical approaches. Moreover, theoretical estimations provide a standardized framework that facilitates comparisons across studies, supports model-based analyses, and allows integration with predictive algorithms such as machine learning.

Table 1. Factors influencing running economy

Parameter	Definition	Relevance in Running Economy
Pace (min/km)	Time taken by the runner to cover one kilometer.	Directly related to running speed and energy cost (VO_2). Higher speeds increase oxygen demand. It is the basis for calculating classic RE formulas [5].
Heart rate (bpm, beats per minute)	Number of heartbeats per minute.	Linearly related to VO_2 during submaximal effort; allows estimation of relative exercise intensity. It is a practical internal load indicator but requires individual calibration [18].
Power (W)	Mechanical work performed per unit of time while running.	Strongly correlated with energy cost and mechanical efficiency. Allows RE estimation without direct VO_2 measurements [2].
Cadence (spm, steps per minute)	Number of steps per minute.	Influences biomechanical efficiency. Optimal cadence reduces vertical oscillations and improves stability, decreasing energy cost [5].
Ground contact time (ms)	Duration for which the foot stays in contact with the ground during each stride.	Shorter contact times are associated with better muscle-tendon elasticity, increasing efficiency and reducing metabolic cost [4].
Stride length (m)	Distance covered in each step while running.	Proper stride length, adjusted to speed and cadence, optimizes energy expenditure. Excessive lengths may increase oxygen cost and injury risk [24].

Building on this foundation, the following paragraph introduces the theoretical formulation commonly used to estimate running economy.

Theoretical Calculation: RE can be estimated using direct laboratory measurements of oxygen consumption or via theoretical formulas relating running velocity and metabolic cost. Eq. 1 shows the standard theoretical expression for RE:

$$\text{RE} = \frac{\text{VO}_2}{v}. \tag{1}$$

Where VO_2 is relative oxygen consumption (ml·kg^{-1}·min^{-1}) and v represents running velocity (km·min^{-1}) [5]. Under level running and steady-state conditions, oxygen consumption can be estimated by Eq: 2.

$$\text{VO}_2 = 3.5 + 0.2\,v. \tag{2}$$

Eq. 3 applies when the incline of the running surface is considered. In this context, the term *grade* represents the slope or incline of the terrain, expressed as a decimal fraction (e.g., a 5% incline corresponds to a grade of 0.05). The model, therefore, incorporates an additional term that accounts for the increased metabolic cost associated with uphill running:

$$\text{VO}_2 = 3.5 + 0.2\,v + 0.9\,v \cdot \text{grade}. \tag{3}$$

These theoretical frameworks enable RE estimation without the need for laboratory equipment, providing scalable tools for performance evaluation and integration with wearable-based monitoring systems.

Wearable Measurement and Validation: Wearable devices, such as smartwatches, foot pods, inertial measurement units (IMUs), and heart rate

monitors, enable the continuous measurement of key running metrics under ecological conditions [32, 18]. While commercially available devices demonstrate strong validity for cadence and stride parameters, others—such as ground reaction forces—may require filtering or advanced modeling to ensure accuracy [10, 27].

Deep learning models further enhance the precision of gait parameter estimation, reducing error relative to laboratory reference systems [31]. These advancements broaden access to performance analytics, allowing everyday athletes to benefit from technology previously reserved for elite research settings [8].

2.3 Deep Learning and Long Short-Term Memory Networks

Deep Learning Foundations: Artificial Intelligence (AI) encompasses methods that enable computational systems to perform cognitive tasks, such as learning and reasoning [6]. Within AI, Machine Learning (ML) focuses on developing algorithms that learn patterns from data, whereas Deep Learning (DL) employs multi-layered neural networks to model complex nonlinear relationships [12]. DL excels in temporal and biological signal processing, making it well-suited for analyzing biomechanical and physiological data [16].

DL reduces dependence on manual feature engineering by learning hierarchical representations, allowing dynamic modeling of training-related adaptations and fatigue patterns in running [4]. As a result, DL facilitates individualized feedback systems that can enhance running performance.

LSTM Models for Running Economy Prediction: RNNs process sequential data by preserving information across time steps; however, traditional RNNs struggle with long-range dependencies due to the vanishing gradient problem [12]. LSTMs overcome this limitation through gated memory mechanisms, enabling effective learning of long-term temporal patterns [13].

LSTMs are particularly effective for modeling dynamic physiological signals such as heart rate, running power, cadence, and contact time, all of which evolve during prolonged running sessions [24]. Attention-enhanced LSTMs further improve performance by allocating computational focus to the most relevant time steps, thereby enhancing prediction accuracy without incurring prohibitive computational costs [24].

LSTM-based approaches demonstrate robustness across diverse runner profiles, lowering the need for highly individualized models and enabling practical application in recreational running contexts [4].

2.4 Model Evaluation

Evaluating deep learning models involves assessing both predictive accuracy and the ability to generalize to unseen data [12]. In the context of RE estimation, model performance was quantified using three complementary statistical metrics. The first metric, the Mean Squared Error (MSE), measures the average squared difference between predicted and actual RE values, providing greater sensitivity to large deviations. It is defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4}$$

A second metric, the Mean Absolute Error (MAE), computes the average absolute difference between predicted and observed values, offering a more robust assessment of typical prediction deviations and being less affected by outliers. It is expressed as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{5}$$

Finally, the Coefficient of Determination (R^2) represents the proportion of variance in RE explained by the model, reflecting its overall predictive fidelity. It is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{6}$$

Here, y_i denotes the actual RE value, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of observed values, and n the number of observations.

To ensure robust evaluation, the model was validated on data from runners excluded from the training set. Additionally, sensitivity analyses were conducted for key input variables, including cadence, heart rate, and pace. These procedures confirm that the predicted RE values provide reliable and actionable information, supporting the design of personalized training programs, optimization of running performance, and reduction of injury risk [4, 24].

3 Models for Running Economy

This section describes the implementation of the three models developed to predict RE based on temporal patterns derived from running activity data. The models include a proposed LSTM model, an alternative LSTM model, and an RNN.

All models were implemented in Python using TensorFlow and Keras as the primary frameworks.

3.1 Dataset and Preprocessing

The dataset used in this study was generated from historical running activity data collected using a Garmin Epix Pro (Gen. 2) smartwatch.

According to the device's technical documentation, the watch captures multiple physiological and biomechanical variables relevant for performance analysis, including average pace, heart rate, running power, cadence, ground contact time, and stride length [11]. These measurements form the core input features used for model training and evaluation. The raw data was exported as a CSV (comma-separated value) file. Table 2 shows a portion of the dataset used for model training.

To ensure data consistency, a preprocessing pipeline was implemented in Python. First, the average pace was converted from `mm:ss` format to minutes per kilometer as a floating-point number. Then, VO_2 (ml/kg/min) and RE (ml/kg/km) were computed for each data segment using the Eq. 1-3 proposed by Barnes and Kilding in 2015 [5]. The relevant features were extracted and standardized using the `StandardScaler` implementation from the `scikit-learn` library, which applies z-score normalization to improve numerical stability during training and ensure that all input variables contribute proportionally to the learning process [20].

Because each model was designed to process sequential data, an additional temporal dimension was added to the feature set. This resulted in an input shape of $(\text{samples}, 1, \text{features})$, where each sample corresponds to a running segment (1 km).

3.2 Proposed Architectures

Three different neural network architectures were developed and evaluated to predict RE from temporal running data. The first model, referred to as the proposed LSTM model, was built using the `Sequential` API from TensorFlow/Keras. Its architecture consists of an LSTM layer with 64 units, followed by a dropout layer with a rate of 0.2 to reduce overfitting. A dense layer with 32 neurons and ReLU activation was added to extract higher-level representations, followed by a final dense layer with linear activation to produce the RE prediction. The model was compiled with the Adam optimizer and Mean Squared Error (MSE) as the loss function.

The second model, the alternative LSTM model, follows a similar structure but with different hyperparameters to explore the impact of model depth and regularization. It includes two stacked LSTM layers with 50 and 25 units, respectively, and a dropout rate of 0.3. This configuration allows the model to capture more complex temporal dependencies at the cost of higher computational load.

The internal structure of the LSTM units used in both architectures is illustrated in Figure 1. Each LSTM cell incorporates a set of gating mechanisms (input, forget, and output gates) that regulate the flow of information across time steps. These gates control how new inputs are integrated, how past information is retained or discarded, and how the hidden state is updated at each timestep. This design enables the network to maintain a cell state that propagates through the sequence, allowing it to capture long-range temporal dependencies that are essential when modeling physiological and biomechanical patterns.

Table 2. Portion of the dataset used for model training

Row	Avg Pace (min/km)	Avg HR (bpm)	Avg Power (W)	Avg Run Cadence (spm)	Avg Ground Contact Time (ms)	Avg Stride Length (m)
1	5:25	144	341	171	257	1.07
2	5:13	154	362	172	252	1.12
3	5:39	147	333	167	257	1.05
⋮	⋮	⋮	⋮	⋮	⋮	⋮
267	5:20	171	361	174	251	1.07
268	5:16	172	361	175	249	1.08
269	5:02	176	376	177	244	1.12

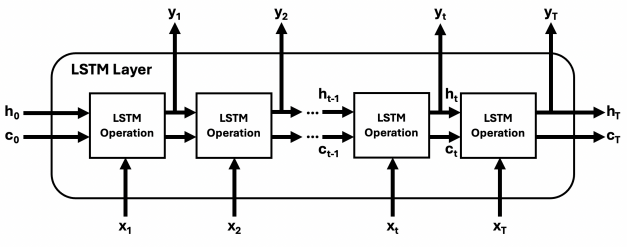

Fig. 1. Architecture of the LSTM layer used in proposed and alternative models

Figure 1 highlights how hidden states ($\mathbf{h_t}$) and cell states ($\mathbf{c_t}$) are passed between consecutive LSTM operations, emphasizing the sequence-to-sequence processing that underpins both LSTM-based models.

In the architecture, $\mathbf{h_t}$ represents the output, and $\mathbf{c_t}$ represents the cell state at time step t. If the layer outputs the entire sequence, it produces $\mathbf{y_1}, \ldots, \mathbf{y_T}$, which corresponds to $\mathbf{h_1}, \ldots, \mathbf{h_T}$. If the layer outputs only the last time step, it produces $\mathbf{y_T}$, which is comparable to $\mathbf{h_T}$. The number of output channels matches the number of hidden units in the LSTM layer.

The third model, the RNN model, was implemented using a simple recurrent layer with 64 units, followed by a dense layer with 32 neurons and ReLU activation, and a final dense layer with linear activation. This model serves as a baseline to compare the performance improvements achieved by the LSTM architectures. All models were trained using the same optimizer (Adam) and loss function (MSE) to ensure consistency across experiments.

3.3 Training and Testing Procedure

To ensure a consistent evaluation framework across all models, the preprocessed dataset was divided into training and testing subsets using an 80/20 split, maintaining the temporal order of the data to preserve its sequential dependencies.

This approach prevents data leakage and ensures that future time steps are never used to predict past values.

Each model, the proposed LSTM model, the alternative LSTM model, and the RNN, was trained under the same experimental conditions to guarantee a fair comparison. Early stopping and model checkpoint callbacks were employed to avoid overfitting and to automatically retain the model weights corresponding to the lowest validation loss. Training was conducted for a maximum of 500 epochs with a batch size of 16, while 15% of the training data was allocated for validation. The models were optimized using the Adam optimizer and the mean squared error (MSE) loss function, with mean absolute error (MAE) and MSE used as performance metrics.

The proposed LSTM model was fine-tuned through iterative experimentation, adjusting the number of LSTM units, dropout rate, and activation functions to optimize generalization and stability in predicting Running Economy. The final selected hyperparameters are summarized in Table 3.

All hyperparameters of the three models were carefully selected through a systematic process of experimentation and validation to ensure comparability and optimize predictive performance. The recurrent layer type was defined according

Table 3. Hyperparameters used in the three models

Hyperparameter	Proposed LSTM model	Alternative LSTM model	RNN model
Recurrent layer type	LSTM	LSTM	SimpleRNN
Number of units	64	100	64
Dropout rate	0.20	0.15	0.20
Hidden dense layer units	32	32	32
Hidden activation function	ReLU	Tanh	ReLU
Output activation function	Linear	Linear	Linear
Optimizer	Adam	Adam	Adam
Loss function	MSE	MSE	MSE
Metrics	MAE, MSE	MAE, MSE	MAE, MSE
Batch size	16	16	16
Validation split	15%	15%	15%
Max epochs	500	500	500
Early stopping	Yes	Yes	Yes
Model checkpoint	Yes	Yes	Yes

to the temporal learning objective: both the proposed and alternative models employed LSTM layers, which are capable of retaining long-term dependencies.

In contrast, the RNN model used a SimpleRNN layer as a baseline to evaluate the effect of removing gating mechanisms. The number of recurrent units was a critical factor; 64 units in the proposed LSTM and RNN models provided an optimal balance between complexity, stability, and computational cost, whereas increasing this number to 100 units in the alternative LSTM slightly improved training accuracy but led to overfitting and reduced generalization. The dropout rate served as a regularization parameter, preventing overfitting by randomly deactivating neurons during training. A rate of 0.20 in the proposed LSTM and RNN models achieved good generalization. In contrast, the lower rate of 0.15 in the alternative LSTM allowed for greater information retention but resulted in higher variance in the validation loss.

All models included a hidden dense layer of 32 neurons to enhance feature transformation before the output stage. The activation function in this hidden layer was set to ReLU for the proposed LSTM and RNN models, promoting faster convergence and avoiding vanishing gradients, while the alternative LSTM used the *tanh* activation to analyze its effect on nonlinearity and smooth transitions between states. The output activation function was linear in all models, consistent with the continuous regression nature of Running Economy estimation. The Adam optimizer was employed in all cases for its adaptive learning rate and robust performance in non-stationary time-series data. The loss function was defined as MSE, which penalizes large prediction deviations more severely, facilitating stable learning dynamics.

The models were trained using a batch size of 16, which provided stable gradient estimation without excessive memory consumption. The training process was limited to a maximum of 500 epochs to ensure convergence. Early stopping and model checkpoint callbacks were implemented to monitor validation loss and automatically store the weights of the best-performing model, thus preventing overfitting and preserving generalization.

A validation split of 15% of the training data was used to evaluate model performance during training without compromising the size of the training subset. The metrics monitored during training were MSE and MAE, which were selected to assess both absolute and squared deviations between the predicted and actual Running Economy values.

Overall, the selection of these hyperparameters was the result of iterative testing and performance benchmarking. The proposed LSTM model demonstrated superior generalization, smoother convergence, and lower validation error compared to the alternative LSTM and RNN models.

These results confirmed that its configuration, 64 LSTM units, 0.20 dropout, ReLU activation, and Adam optimization, offered the most robust and efficient balance between prediction accuracy and computational efficiency for estimating RE.

3.4 Feedback, Results, and Metrics

After training, the three models were evaluated on the test set to assess their predictive performance in estimating RE. For consistency, all models were evaluated using three standard regression metrics: MSE, MAE, and R^2. These metrics provided a quantitative basis for comparing model accuracy and generalization.

Additionally, model performance was analyzed through two main types of visualizations. The first compared the predicted and actual RE values for both training and testing sets, highlighting how closely each model replicated real data patterns.

The second visualization displayed the evolution of the loss metrics (MSE and MAE) over training epochs, offering insight into model convergence and overfitting tendencies. These analyses collectively allowed for an in-depth comparison of the three architectures and the identification of the most effective approach for RE estimation.

4 Results

This section presents the results obtained from the experimental evaluation of the proposed approach. The findings provide an overview of the models performance and serve as the basis for the analysis and discussion presented in the following sections.

4.1 Experiments Configuration

The LSTM and RNN models were trained to estimate the RE. The complete dataset was divided into training (80%) and testing (20%) subsets, preserving the temporal sequence of the running session. All input and target variables were standardized before model training.

The experiments were performed on an Apple MacBook Air (M2, 2022) equipped with an Apple M2 system-on-a-chip (SoC) that integrates an 8-core CPU (4 performance cores and 4 efficiency cores) with a base frequency of 3.49 GHz, and 8.00 GB of unified memory. The operating system used was macOS Monterey 12.5. The GPU employed is an integrated Apple M2 GPU with 8 cores.

4.2 Models Training Performance

All three models were trained for a maximum of 500 epochs using the Adam optimizer and MSE as the loss function. An early stopping criterion with a patience of 20 epochs was applied, monitoring the validation MSE to prevent overfitting and ensure efficient training. Each model incorporated dropout regularization and dense layers to balance complexity and generalization.

The proposed LSTM model consisted of a single LSTM layer with 64 units, a dropout rate of 0.2, and two dense layers with 32 and 1 neurons, respectively. The alternative LSTM model included two stacked LSTM layers with 50 and 25 units and a higher dropout rate of 0.3 to test deeper temporal representations. Finally, the RNN model employed a SimpleRNN layer with 64 units, followed by a dropout rate of 0.2 and the same dense layer configuration used in the LSTM models.

As illustrated in Figures 2, 4, and 6, all models exhibited stable convergence with decreasing training and validation losses over time. The proposed LSTM model demonstrated the most efficient and consistent learning curve, with the validation MSE plateauing after approximately 100 epochs, which triggered early stopping and effectively prevented overfitting.

In contrast, the alternative LSTM model exhibited a slightly slower convergence trend due to its deeper structure and higher dropout rate, whereas

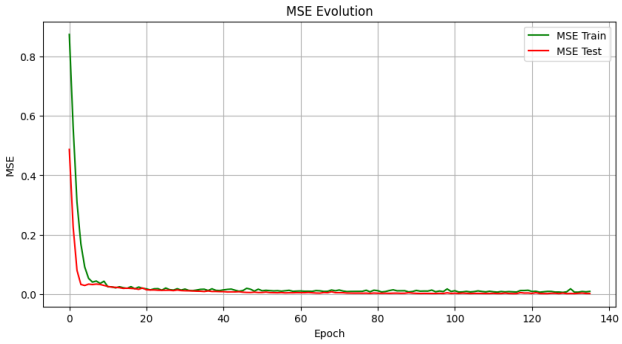

Fig. 2. Evolution of the MSE during model training and validation for the proposed LSTM model

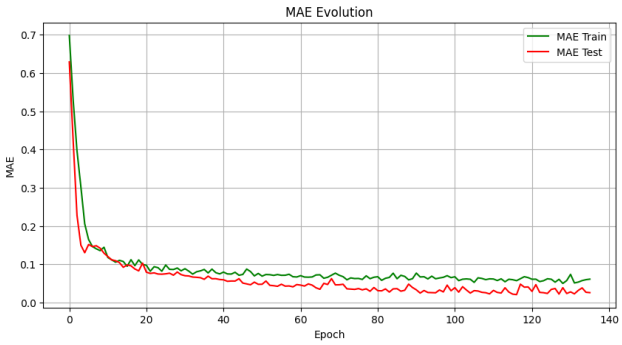

Fig. 3. Evolution of the MAE during model training and validation for the proposed LSTM model

the RNN model displayed greater oscillation in the validation loss, indicating a reduced ability to capture long-term temporal dependencies.

Similarly, the evolution of the MAE shown in Figures 3, 5, and 8 revealed comparable trends across models. The proposed LSTM model consistently achieved the lowest MAE values with minimal variance across epochs, confirming its superior stability and predictive precision.

The alternative LSTM model presented slightly higher MAE values but maintained a smooth convergence profile, while the RNN model exhibited greater fluctuation and slower error reduction. These results highlight the robustness and reliability of the proposed LSTM architecture in learning complex temporal patterns associated with RE.

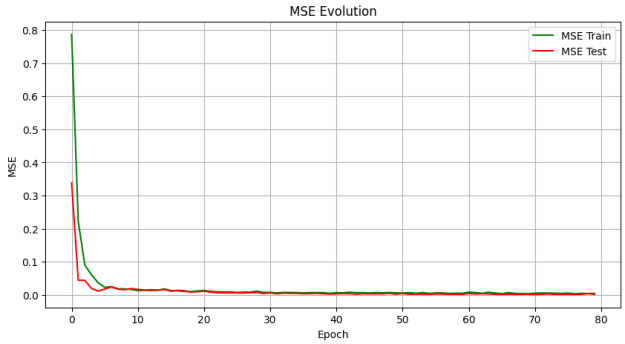

Fig. 4. Evolution of the MSE during model training and validation for the alternative LSTM model

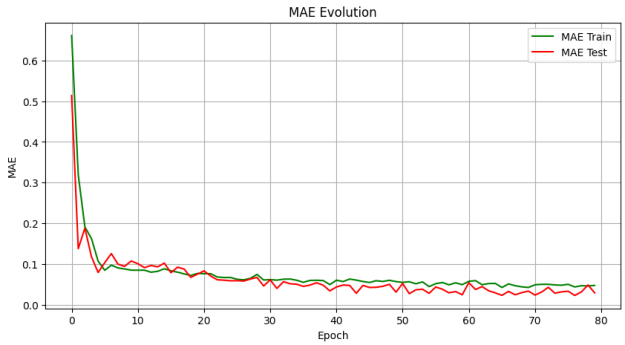

Fig. 5. Evolution of the MAE during model training and validation for the alternative LSTM model

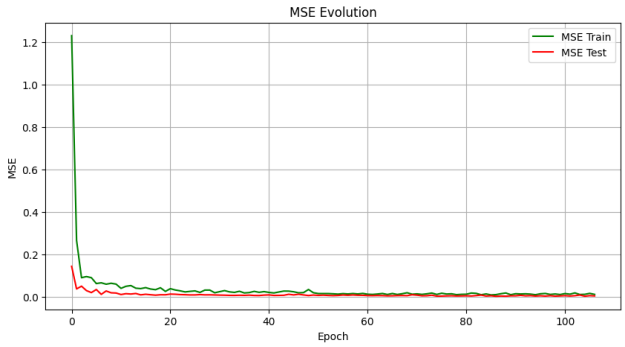

Fig. 6. Evolution of the MSE during model training and validation for the RNN model

4.3 Predictive Accuracy

After training, all three models were evaluated on the unseen test set to assess their predictive

performance in estimating RE. Table 4 summarizes the results obtained for each architecture based on the MSE, MAE, and R^2.

The results indicate that all models achieved strong predictive performance, with R^2 values exceeding 0.97, demonstrating their ability to capture the temporal dynamics underlying RE.

Among them, the proposed LSTM model achieved the best overall performance, yielding the lowest MSE (0.036995) and a high R^2 value of 0.986986, which explains approximately 98.7% of the variance in the test data. The alternative LSTM model achieved comparable results, with slightly higher MSE but marginally lower MAE, suggesting effective yet less consistent generalization compared to the proposed configuration. The RNN model, while still performing well, exhibited higher prediction errors and lower correlation with the ground truth, reinforcing the advantage of LSTM architectures in learning long-term temporal dependencies.

4.4 Temporal Prediction of Running Economy

Figures 7, 9, and 10 illustrate the comparison between the actual and predicted RE values across the complete time series, including both training and testing segments. For the proposed LSTM model (Figure 7), the predicted values closely follow the real data, with the training predictions (green line) showing excellent alignment and the test predictions (red line) maintaining the overall trend with minimal dispersion. The alternative LSTM model (Figure 9) also demonstrates a strong correspondence between predicted and observed RE values, though with slightly higher variability in the testing interval, likely due to its deeper architecture and stronger regularization. In contrast, the RNN model (Figure 10) presents a less precise match, with wider deviations from the true RE values, particularly during rapid temporal transitions.

In all cases, the vertical dashed line marks the division between training and testing intervals. The visual comparison confirms that the LSTM-based architectures, particularly the proposed LSTM model, effectively capture the temporal dependencies embedded in the physiological and biomechanical signals, enabling accurate estimation of RE throughout the time series.

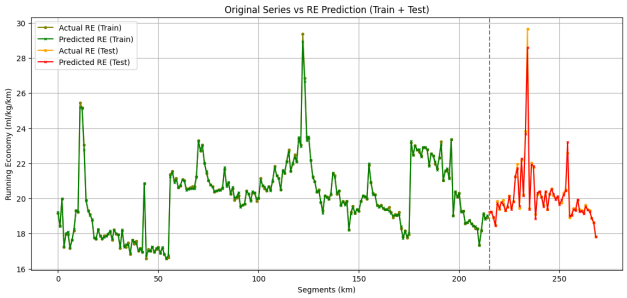

Fig. 7. Comparison between actual and predicted RE across training and testing segments for the proposed LSTM model. The dashed line marks the transition from training to testing data

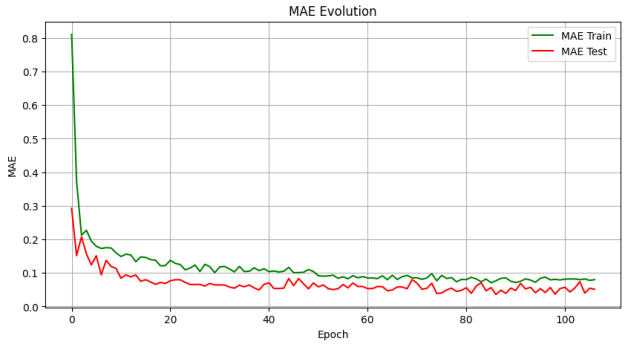

Fig. 8. Evolution of the MAE during model training and validation for the RNN model

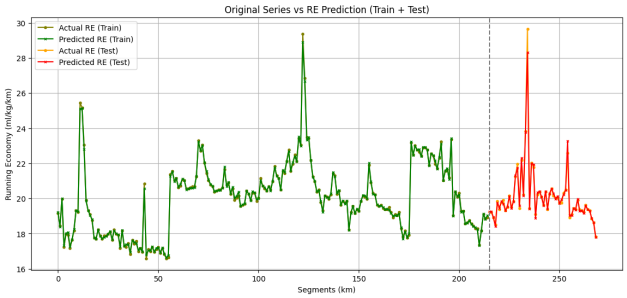

Fig. 9. Comparison between actual and predicted RE across training and testing segments for the alternative LSTM model. The dashed line marks the transition from training to testing data

Table 4. Performance comparison between RNN and LSTM models

Model	Test MSE	Test MAE	Test R^2
LSTM (proposed)	0.036995	0.102474	0.986986
LSTM (alternative)	0.048944	0.098615	0.982782
RNN	0.080312	0.141323	0.971748

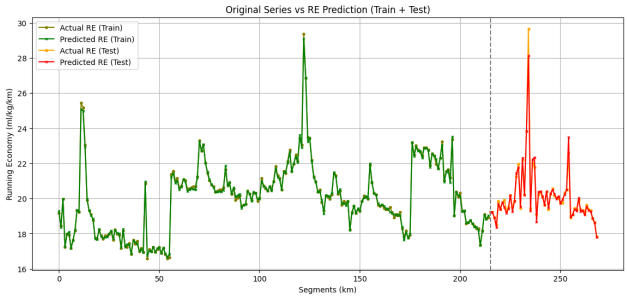

Fig. 10. Comparison between actual and predicted RE across training and testing segments for the RNN model. The dashed line marks the transition from training to testing data

4.5 Discussion of Models Behavior

All three models were able to learn meaningful relationships between the biomechanical and physiological variables used as input and the resulting RE. The LSTM-based architectures demonstrated a superior ability to capture complex, nonlinear dependencies within the sequential data compared to the RNN model. Despite using a single-timestep input representation (per-kilometer data), both LSTM models effectively preserved temporal continuity and achieved strong generalization on unseen data, validating their capacity to extract dynamic patterns relevant to RE estimation.

The proposed LSTM model exhibited the most stable and consistent predictive behavior, maintaining low error metrics and smooth convergence trends across epochs. Its moderate number of units and dropout rate allowed it to balance model complexity and regularization effectively. The alternative LSTM model, while achieving comparable accuracy, showed slightly slower convergence and greater sensitivity to regularization due to its deeper architecture. In contrast, the RNN model presented higher variability in predictions and weaker alignment with real data during testing, reflecting its limited ability to retain long-term temporal dependencies inherent in running dynamics.

Overall, the comparative analysis confirms that deep recurrent architectures, particularly the LSTM networks, are well-suited for modeling RE using wearable sensor data. The strong predictive performance of the proposed LSTM model highlights its potential as a practical, non-invasive alternative to traditional laboratory-based metabolic assessments, offering a viable solution for continuous, real-world monitoring of running efficiency.

5 Conclusions

This work presents a data-driven approach for estimating RE using wearable-derived biomechanical and physiological metrics, along with recurrent neural network architectures. By leveraging temporal patterns from real running sessions, the study explored and compared three different deep learning models: the proposed LSTM model, the alternative LSTM model, and a baseline RNN model. The objective was to determine the most effective architecture for predicting RE in real-world conditions without the need for specialized laboratory testing.

The experimental results demonstrated that all three models were capable of learning meaningful relationships between input metrics such as pace, heart rate, power, cadence, ground contact time, and stride length, and the resulting RE values.

However, clear performance differences emerged. The proposed LSTM model consistently achieved the lowest prediction errors, the highest coefficient of determination, and the most stable convergence behavior across training and validation epochs. Its balanced architecture,

moderate number of LSTM units, and appropriate regularization allowed it to generalize effectively while preserving temporal dependencies in the data. The alternative LSTM model demonstrated competitive performance but exhibited slower convergence and slightly higher variability. In contrast, the RNN model underperformed, confirming its limitations in modeling the long-term dependencies inherent in running dynamics.

The results of this work validate the feasibility of estimating RE using deep learning and wearable technology. The ability to accurately predict RE without laboratory equipment represents a significant advancement, offering athletes and coaches a non-invasive, accessible, and continuous method for monitoring running efficiency. This approach has the potential to support personalized training, optimize performance strategies, and extend the utility of consumer-grade devices in endurance sports analytics.

Future work could extend this research by integrating longer temporal windows, experimenting with transformer-based architectures, expanding the dataset across multiple athletes, and validating predictions against laboratory-measured RE to further improve robustness and generalizability. The findings presented here establish a strong foundation for developing intelligent, field-deployable tools for endurance performance assessment.

References

1. **Advaitha Vetagiri, P. P., Prateek Mogha**, **.** Multilate classifier: A novel ensemble of CNN-BiLSTM with ResNet-based Multimodal Classifier for AI-generated Hate Speech Detection, journal = Computación y Sistemas, year = 2025, volume = 29, number = 3, pages = , publisher = IPN, note = ISSN 1405-5546. .
2. **Austin, C. L., Hokanson, J. F., McGinnis, P. M., Patrick, S. (2018).** The relationship between running power and running economy in well-trained distance runners. Sports, Vol. 6, No. 4, pp. 142. DOI: 10.3390/sports6040142.
3. **Babakhani, S., Remy, D., Roitberg, A. (2025).** Deep learning for metabolic rate estimation from biosignals: A comparative study of architectures and signal selection. arXiv preprint.
4. **Bailey, C. A., Mir-Orefice, A., Uchida, T. K., Riazati, H., Delp, S. L. (2023).** Smartwatch-based prediction of single-stride and stride-to-stride gait outcomes using regression-based machine learning. Annals of Biomedical Engineering, Vol. 51, pp. 2504–2517. DOI: 10.1007/s10439-023-03290-2.
5. **Barnes, K. R., Kilding, A. E. (2015).** Running economy: measurement, norms, and determining factors. Sports Medicine - Open, Vol. 1, No. 1, pp. 8. DOI: 10.1186/s40798-015-0007-y.
6. **Borgia, B., Dufek, J. S., Silvernail, J. F., Radzak, K. N. (2022).** The effect of fatigue on running mechanics in older and younger runners. Gait & Posture, Vol. 97, pp. 86–93. DOI: 10.1016/j.gaitpost.2022.07.249.
7. **Briand, J., di Prampero, P. E., Osgnach, C., Thibault, G., Tremblay, J. (2025).** Quantifying metabolic energy contributions in sprint running: a novel bioenergetic model. European Journal of Applied Physiology. DOI: 10.1007/s00421-025-05831-0.
8. **Carrier, B., Navalta, J. W. (2022).** Data analysis processes and techniques for validation of wearable technology: An example. Topics in Exercise Science and Kinesiology, Vol. 3, No. 1, pp. Article 10.
9. **Colunga-Rodriguez, A. A., Martínez-Rebollar, A., Estrada-Esquivel, H., Clemente, E. (2025).** Modelo híbrido de programación genética y redes neuronales para el reconocimiento de emociones. Computación y Sistemas, Vol. 29, No. 3. ISSN 1405-5546.
10. **de Fontenay, B. P., Roy, J. S., Dubois, B., Bouyer, L., Esculier, J. F. (2020).** Validating commercial wearable sensors for running gait parameters estimation. IEEE Sensors

Journal, Vol. 20, No. 14, pp. 7783–7791. DOI: 10.1109/JSEN.2020.2982568.

11. **Garmin Ltd. (2023).** Epix Pro (Gen 2) — Owner's Manual. Garmin. Available from Garmin Support Documentation.
12. **Goodfellow, I., Bengio, Y., Courville, A. (2016).** Deep Learning. MIT Press. DOI: 10.7551/mitpress/10993.001.0001.
13. **Hochreiter, S., Schmidhuber, J. (1997).** Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
14. **Hsiao, C. T., Tong, C., Coté, G. L. (2025).** Machine learning-based vo2 estimation using a wearable multiwavelength photoplethysmography device. Biosensors, Vol. 15, No. 4, pp. 208. DOI: 10.3390/bios15040208.
15. **Jaszczak, B., Plociniczak, L. (2024).** Optimal strategy for trail running with nutrition and fatigue factors.
16. **LeCun, Y., Bengio, Y., Hinton, G. (2015).** Deep learning. Nature, Vol. 521, pp. 436–444. DOI: 10.1038/nature14539.
17. **Lovell, D. I., Stuelcken, M., Eagles, A. (2025).** Exercise testing for metabolic flexibility: Time for protocol standardization. Sports Medicine - Open, Vol. 11, No. 1, pp. 31.
18. **Parak, J., Uuskoski, M., Machek, J., Korhonen, I. (2017).** Estimating heart rate, energy expenditure, and physical performance with a wrist photoplethysmographic device during running. JMIR mHealth and uHealth, Vol. 5, No. 7, pp. e97. DOI: 10.2196/mhealth.7437.
19. **Paraschiakos, S., de Sá, C. R., Okai, J., Slagboom, P. E., Beekman, M., Knobbe, A. (2022).** A recurrent neural network architecture to model physical activity energy expenditure in older people. Data Mining and Knowledge Discovery, Vol. 36, No. 1, pp. 477–512.
20. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011).** Scikit-learn: Machine learning in Python. J. Mach. Learn. Res., Vol. 12, pp. 2825–2830.
21. **Qureshi, T. S., Shahid, M. H., Farhan, A. A., Alamri, S. (2025).** A systematic literature review on human activity recognition using smart devices: advances, challenges, and future directions. Artif. Intell. Rev., Vol. 58, No. 9.
22. **Saad, H. S., Zaki, J. F. W., Abdelsalam, M. M. (2024).** Employing of machine learning and wearable devices in healthcare system: tasks and challenges. Neural Computing and Applications, Vol. 36, pp. 17829–17849. DOI: 10.1007/s00521-024-10197-z.
23. **Saunders, P., Pyne, D., Telford, R., Hawley, J. (2004).** Factors affecting running economy in trained distance runners. Sports Medicine, Vol. 34, No. 7, pp. 465–485. DOI: 10.2165/00007256-200434070-00005.
24. **Shao, Y., Li, R.-D., Luo, Y.-J., Zhu, M. (2021).** Research on running data analysis method based on attention-lstm. Proceedings of the 2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE, pp. 446–450. DOI: 10.1109/ICITBS53129.2021.00116.
25. **Song, Y., Wang, Z., Wang, H., Sun, G., Gong, B., Zhang, F. (2025).** Interpretable deep learning for personalized energy expenditure prediction using ecg and acceleration signals in incremental exercise. Scientific Reports, Vol. 15, No. 1, pp. 36277.
26. **Strohrmann, C., Harms, H., Tröster, G. (2012).** Monitoring fatigue in running using body-worn sensors. 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, IEEE, pp. fatigue–monitoring–1–6. DOI: 10.1109/BSN.2012.30.
27. **Ursul, I., Pereymybida, A. (2023).** Unsupervised detection of anomalous running patterns using cluster analysis. Proceedings of the 2023 IEEE 13th International Conference on Electronics and Information

Technologies (ELIT), IEEE, pp. 148–152. DOI: 10.1109/ELIT61488.2023.10310751.

28. **Van Hooren, B., Jukic, I., Cox, M., Frenken, K. G., Bautista, I., Moore, I. S. (2024).** The relationship between running biomechanics and running economy: A systematic review and meta-analysis of observational studies. Sports Medicine, Vol. 54, No. 5, pp. 1269–1316. DOI: 10.1007/s40279-024-01997-3.
29. **Wang, Z., Song, Y., Pang, L., Li, S., Sun, G. (2025).** Attention-enhanced CNN-LSTM model for exercise oxygen consumption prediction with multi-source temporal features. Sensors, Vol. 25, No. 13, pp. 4062. DOI: 10.3390/s25134062.
30. **Wenzel, C., Liebig, T., Swoboda, A., Smolareck, R., Schlagheck, M. L., Walzik, D., Groll, A., Goulding, R. P., Zimmer, P. (2024).** Machine learning predicts peak oxygen uptake and peak power output for customizing cardiopulmonary exercise testing using non-exercise features. European Journal of Applied Physiology, Vol. 124, No. 11, pp. 3421–3431.
31. **Zhang, H., Guo, Y., Zanotto, D. (2020).** Accurate ambulatory gait analysis in walking and running using machine learning models. IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 28, No. 1, pp. 191–202. DOI: 10.1109/TNSRE.2019.2958679.
32. **Zhang, T., Zhang, M., Luo, J. (2025).** The application and development trends of wearable devices (wd) in endurance sports training: A literature review. Quality in Sport, Vol. 37, pp. 57604. DOI: 10.12775/QS.2025.37.57604.

Article received on 31/10/2025; accepted on 15/12/2025.
**Corresponding author is Ulises Orozco-Rosas.*