Pilot Study on Predicting Maximal Oxygen Uptake Using LSTM Models and Physiological Signals

Jaiveer Bagga

Received May 6, 2025
Accepted September 25, 2025
Electronic access November 15, 2025

Maximal oxygen consumption (VO_2max) strongly indicates an individual's fitness level. However, traditional methods of measuring VO_2max often require expensive equipment and controlled laboratory settings. This study explores the feasibility of estimating VO_2max using physiological signals collected from seven subjects on a cycle ergometer. Data from multiple cycling protocols were processed to extract features, including power output, heart rate, cadence, and respiratory frequency. Average oxygen uptake over 20 seconds at peak power (AO20) was used to approximate VO_2max and served as the primary target variable. A long short-term memory (LSTM) neural network was trained on these inputs, achieving a root mean square error (RMSE) of 0.1745 mL/Min and an R^2 score of 0.9693 on the test dataset. While the results indicate that the model can capture key relationships between cycling performance and aerobic capacity, the findings are limited by the small sample size and the controlled ergometer environment. These constraints mean further research with larger and more diverse datasets is required before real-world applicability can be fully assessed.

Keywords: Deep-learning, LSTM, Time-series Prediction, Physiological Signals, VO_2max

Introduction

Maximal oxygen consumption (VO_2max) represents the maximum amount of oxygen the body can utilize during intense exercise and is considered the standard for assessing aerobic fitness in an individual[1,2]. Accurate measurements of VO_2max typically require specialized gas analysis equipment and controlled laboratory protocols, which may not always be available or practical to many[3]. Recent research has begun to explore whether wearable devices and/or machine learning can provide alternatives to lab-based testing; however, these approaches are still relatively new[4–6].

A common approach to predicting VO_2max is through computational models created from real-world data. For example, a 2008 review by Akalan et al. uses VO_2max prediction equations and compares them to numerous data collected across various activities, including walking, running, and cycling[7]. However, equations cannot account for the variability from subject to subject due to differences in populations, thus struggling to estimate the subject's VO_2max accurately[1]. In this scenario, a machine-learning algorithm can process the subject data and account for more variability. Another study, conducted by Borror et al., utilizes an Artificial Neural Network (ANN), more specifically a feed-forward neural network (FFNN), to predict an individual's VO_2[8]. An FFNN processes inputs independently with no memory of past inputs[9]. However, in VO_2max prediction, past data can be valuable because VO_2max levels respond with a slight delay to changes in cardiac and muscular output[10]. This time-sensitive variability suggests that incorporating prior time-series data can improve prediction accuracy. For this reason, this study utilized a Recurrent Neural Network (RNN), as RNNs maintain a hidden state to recall past data and create patterns to predict VO_2max readings accurately[11].

More advanced architectures, such as the Attention-Enhanced CNN-LSTM model proposed by Wang et al. (2025), capture both temporal and spatial patterns in physiological data, which improves accuracy over traditional RNNs[12]. However, this approach relies heavily on large, curated datasets and primarily focuses on exercise metrics obtainable only using lab equipment. Similarly, Hsiao et al. (2025) demonstrated that a wrist-worn multiwavelength photoplethysmography (PPG) device combined with a machine learning model can estimate VO_2 continuously[13]. However, their model is constrained by reliance on a single sensor modality, potentially missing physiological signals that could enhance prediction accuracy. In contrast, the present study addresses these gaps by integrating multiple physiological inputs, including heart rate and respiratory frequency, and using a smaller dataset within a sequential RNN-based framework.

This project aims to develop a machine learning model capable of predicting VO_2max from physiological signals collected during cycling ergometer tests using the average oxygen uptake over 20 seconds at peak power (AO20), an approximate surrogate for VO_2max used in the original paper. Short-duration rolling averages are a recognized method for reducing variability in VO_2 measurements. A 20-second average has been reported as one accepted approach among exercise physiologists, alongside longer averages, such as 30s or 60s, which can correct up to 96-98% of variability[14]. However, AO20 should be regarded as an approximation rather than a direct measurement of VO_2max. The goal is to provide an accurate and efficient alternative to direct VO_2max testing. A Long Short-Term Memory (LSTM) neural network[15,16] was used to take advantage of the flat regions of peak power output. These regions exhibit consistent physiological trends over time when the subject exerts maximum power, which benefits the LSTM, as it is designed to analyze sequential data. By modeling how the features evolve during the test, it was hypothesized that the LSTM can learn patterns to predict an individual's VO_2max without needing oxygen consumption data.

This study demonstrates a proof-of-concept showing that VO_2max can be estimated from easily measurable signals (power, cadence, heart rate, respiratory frequency) obtained from wearable sensors instead of specialized gas equipment. This novel contribution addresses the unmet need for a practical, accessible, and scalable method to estimate aerobic fitness in real-world settings, without the cost, expertise, and equipment burden associated with laboratory-based cardiopulmonary testing.

Methods

Dataset

The data used in this study were obtained from a publicly available, anonymized dataset from Kaggle to ensure that no personal information was released. The dataset was uploaded to Kaggle by the first author of the original peer-reviewed publication from which the data originates, and the additional associated supplementary PDF confirms the experimental methods[17]. All subjects provided informed consent, and the experiment followed the guidelines of the Declaration of Helsinki[18]. The dataset consisted of physiological data sampled at a frequency of 1 Hz from seven subjects (six male, one female) each undergoing the same set of ergometric cycling tests, where power increased over time until the subject reached maximal effort. The subjects had a mean age of 32.69 years and a standard deviation of 10.56 years[19].

Each subject's data was labeled with their measured VO_2, determined from oxygen data, which was then translated into their VO_2max using a rolling average method. Respiratory measurements, including respiratory frequency and VO_2, were measured using a breath-by-breath method from an automated open-circuit gas analyzer (Quark CPET by Cosmed), ensuring measurement reliability[19]. The features used for prediction include power (in watts), cadence (in revolutions per minute), heart rate (in beats per minute), and respiratory frequency (in breaths per minute), all of which are commonly collected during fitness exercise tests.

The problem was treated as a supervised learning regression task using time-series inputs and a continuous target output–VO_2max, which can be approximated using 20-second rolling averages of the subject's VO_2 during peak power output. This method helps to reduce the variability in VO_2 readings[14].

For each subject, the periods of maximal power output were identified (Figure 2, red bounding boxes) and used to calculate AO20 values by averaging the 20-second chunks of oxygen consumption data. A period of maximal power output can yield multiple time series of VO_2 data and their corresponding averages (target values) via a 20-second moving window with one-second increments. This ensures that the model can learn to predict AO20 based on the physiological state during maximal effort. This yielded 4616 time series samples, each 20 seconds long.

The Wingate Anaerobic Test (WAnT) is a 30-second all-out cycle ergometer test used to evaluate anaerobic physiological responses[20,21]. The data from the WAnT (Figure 1) in this study provided a singular peak rather than constant, maximum power regions. This led to the test being excluded, as it was unsuitable for the model's use.

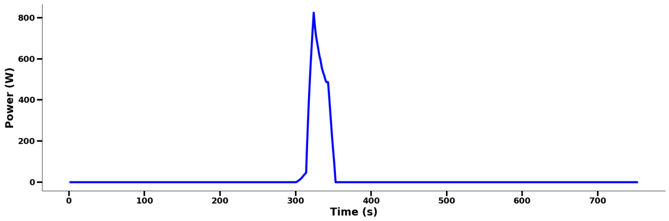

Fig. 1 An example of the power output during the Wingate Anaerobic Test. The peak power output is only over a narrow duration and is unsuitable for use as training data for the RNN model.

Next, the heart rate and respiratory frequency features underwent standard scaling, transforming the data to have a mean of zero and a standard deviation of one across all subjects using the equation:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Where x is the raw value, μ is the mean, σ is the standard deviation, and z is the resulting standardized z-score. This enabled the model to learn physiological patterns more effectively. Power and Cadence were left in their original scales due to their relatively low and stable range of values. The data was split into 60% for training, 20% for validation, and 20% for testing. The split was performed at the sample (segment) level rather than the subject level, meaning that data from all individuals could appear in each partition. While this approach helped maximize the diversity of physiological states the model was exposed to,

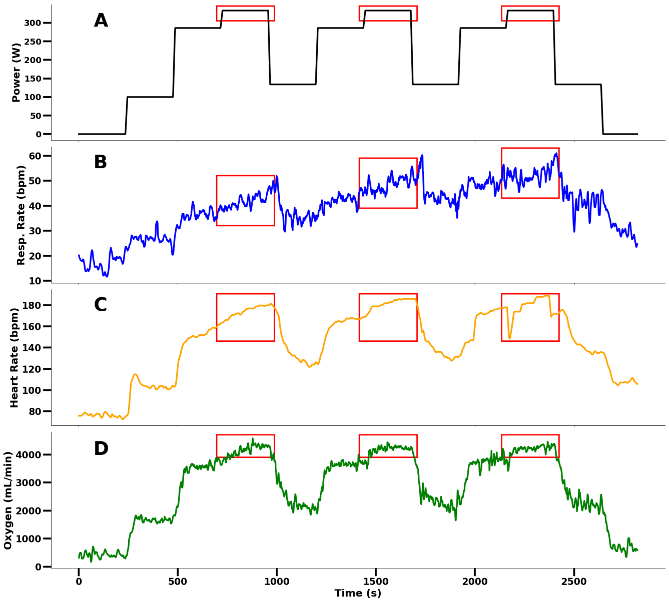

Fig. 2 Time series signals recorded from a subject during an incremental exercise protocol. The maximum power output regions are enclosed within the red bounding boxes in Figure 2A, with corresponding areas highlighted in Figures 2B, 2C, and 2D.

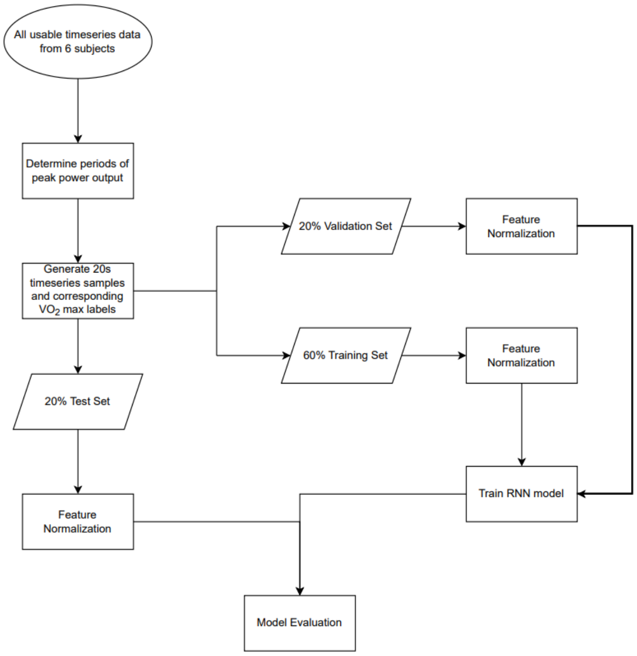
All usable timeseries data from 6 subjects
Determine periods of peak power output
Generate 20s timeseries samples and corresponding VO$_2$ max labels
20% Validation Set
Feature Normalization
60% Training Set
Feature Normalization
20% Test Set
Train RNN model
Feature Normalization
Model Evaluation

Fig. 3 Flowchart representing data processing, model training, and model evaluation.

it does pose the risk of data leakage, as the model may have encountered similar subject-specific patterns across windows. However, this design was intentional for two reasons: (1) exposing the model to a larger distribution of physiological states during training, helping improve its generalization ability, given this small dataset, and (2) evaluating how well the model can predict outcomes on new recordings from the same individual, reflecting real-world applications such as continuous or repeated monitoring of subjects. While subject-level splitting would be required to assess strict cross-subject generalization, the focus of this study was on within-subject predictive performance.

The input to the model is a tensor composed of four physiological features recorded over a 20-second window (heart rate, respiratory frequency, power, and cadence), and the model outputs the VO$_2$max (AO20) prediction for the next second.

The data processing, model training, and evaluation workflow are shown in Figure 3.

Model Architecture, Training, and Evaluation

An LSTM network was employed to predict the VO$_2$max of the subjects. LSTMs are particularly well-suited for handling time-series data due to their ability to preserve temporal dependencies and remember long-term relationships across sequential inputs[15,22], which is essential for modeling physiological processes that change over time. Compared to Gated Recurrent Units (GRUs), LSTMs provide stronger performance on small to medium datasets. Their additional gating mechanisms allow more effective learning of complex patterns, making them a stronger proof-of-concept model choice. While transformers represent a more advanced alternative for sequential modeling, their performance advantage depends more heavily on large datasets[23], which were not available for this study; therefore, using transformers would likely result in overfitting.

The deliberate choice of LSTMs balances its predictive power with a smaller dataset, making it appropriate for this application.

The model architecture consisted of two LSTM layers, each with a hidden vector size of 64. The output of the LSTM layer underwent layer normalization and was then passed through a fully connected layer to produce a single continuous output corresponding to the predicted AO20. The root mean square error (RMSE) and mean absolute percentage error (MAPE) were used to evaluate how well the model was performing, which are given as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \qquad \mathrm{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2)$$

where $\hat{y}_i$ is the predicted VO$_2$max, y_i is the actual VO$_2$max, and n is the total number of data points. The Adam optimizer[24] was used to train the model with a learning rate of 0.001. Training was performed over 68 epochs, and convergence was determined via early stopping with a patience of five. A batch size of eight was used for training. All hyperparameters were optimized through hyperparameter tuning, including learning rate, hidden size, and the number of LSTM layers. A random search strategy was employed across a predefined search space (Table 1), with five-fold cross-validation used to evaluate each candidate configuration. RMSE on the validation folds was used as the optimization metric. Table 2 highlights the parameters and optimized hyperparameter values used in the study.

Table 1 Hyperparameter search space explored during random search optimization.

Hyperparameter	Search Space Values
Learning rate	0.1, 0.01, 0.001, 0.0001
Hidden size	32, 64, 128
Number of LSTM layers	1, 2, 3
Batch size	8, 16, 32
Optimizer	Adam, RMSProp

Table 2 Summary of parameters and hyperparameters used in training the model. Through tuning and optimization, these were determined to be the optimal values.

Parameter / Hyperparameters	Value
Input Size	4
Hidden Size	64
Number of LSTM Layers	2
Output Size	1
Learning Rate	0.001
Batch Size	8

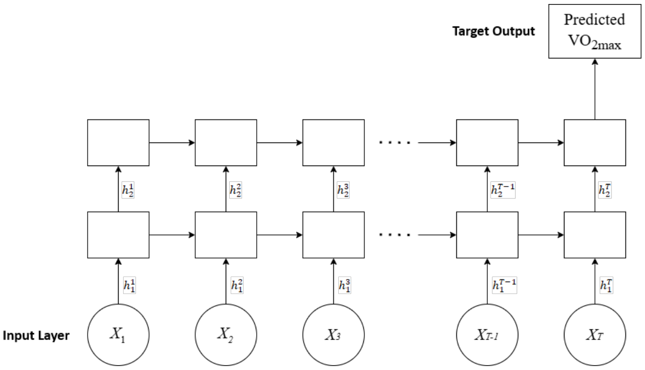

Fig. 4 The many-to-one, two-layer LSTM model architecture used in this study. X_T represents the input sequence at time T, and h_T^L denotes the hidden state of layer L at time T. The final hidden state h_T^2 is fed into a fully connected layer to predict the VO_2max.

The R^2 value and the root mean square error were used to evaluate and determine the model's accuracy. After training and during testing, the data were inversely scaled to restore the features to their original values. The source code and dataset used in this study are available on GitHub.

Results

The trained LSTM model showed encouraging predictive trends, with Figure 5 presenting the training and validation RMSE curves plotted against epochs, Figure 6 exhibiting the scatterplot comparing the predicted and actual VO_2max values, and Figure 8 showing the corresponding residual plot to Figure 6.

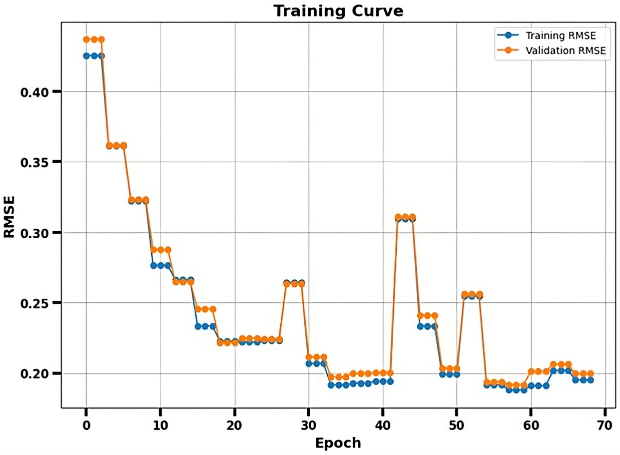

Fig. 5 Training curve (Blue) and validation curve (Orange) are plotted on the RMSE vs. Epoch graph.

The training and validation RMSE curves are highlighted in Figure 5 and show the model's RMSE during training over 68 epochs. The curves generally show a downward trend across epochs; however, RMSE spikes are observed. These fluctuations are expected given the small dataset size, the use of window-splits, and stochastic mini-batch training. With limited training samples, including a batch with harder-to-predict windows may temporarily increase error. Such spikes do not indicate model collapse, as the overall trajectory of the curve appears downwards.

The scatter plot (Figure 6) demonstrates that the model was strong in estimating VO_2max, achieving an R^2 of 0.9693, a MAPE of 2.47%, and an RMSE of 0.1961 mL/min. It shows that most predicted VO_2max values fall close to the actual VO_2max values as they cluster closely around the line. However, the model cannot accurately predict some data points inside the red circle. The R^2 value indicates that approximately 97% of the variance in VO_2max is accounted for by the LSTM regression model, with the remaining 3% because of factors outside the model's scope. To better understand the flattening of the data above 5500 mL/min, Figure 7 presents a histogram revealing a minimal number of samples above this range, which likely caused the model to struggle generalizing beyond this threshold, leading it to predict a horizontal line of VO_2max rather than capturing its variability.

The residuals plotted in Figure 7 correspond to the scatterplot in Figure 6, displaying a generally random distribution around

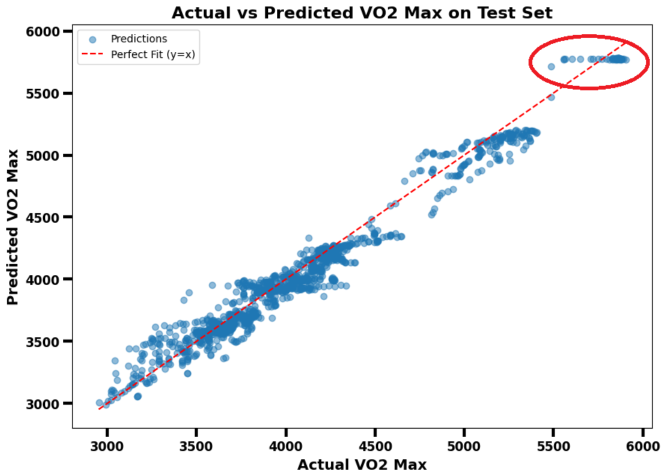

Fig. 6 Predicted vs. actual VO_2max values. The red dashed line denotes a perfect fit. The values in the red circle reveal slight inaccuracies in the model's prediction capabilities.

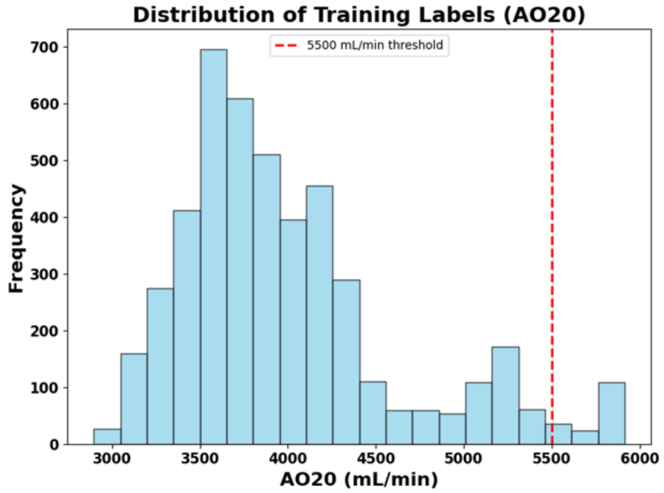

Fig. 7 Histogram of training set AO20 labels. The dashed line marks 5500 mL/min, highlighting the data scarcity above this range.

zero mL/min, except for the outlying points previously highlighted in the red circle. This suggests the model has a relatively low systematic bias when predicting VO_2max.

To contextualize the LSTM's results, a simple linear regression baseline was trained on the same dataset. The model achieved an R^2 of 0.8621, an RMSE of 0.3725 mL/min, and a MAPE of 4.68% on the test set. These results suggest that the LSTM model provided more accurate and reliable VO_2max estimations relative to the actual values compared to the baseline linear regression model.

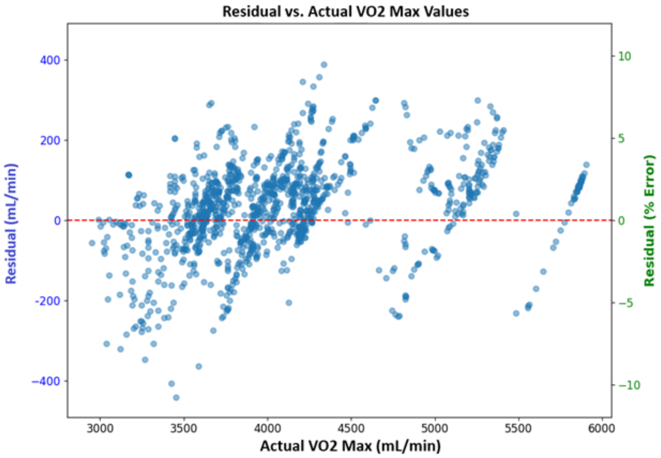

Fig. 8 Residuals (actual minus predicted) vs. predicted VO_2max values. The left axis measures the residuals in mL/min, while the right axis measures the residuals as a percentage of the actual VO_2max values. The red dashed line denotes zero error.

Discussion

The results attained by the model suggest that temporal patterns in the four physiological features—Power, Cadence, Heart Rate, and Respiratory Frequency—are highly informative for predicting VO_2max. The scatter plot (Figure 6) of predicted vs. actual AO20 values shows a strong linear relationship. The residuals vs. predicted values plot (Figure 8) shows patterns of left-to-right displays of residuals about the zero-error line (red dashed line). This indicates that the model can capture the mean VO_2max values for each period of peak power output, but cannot always predict the variation precisely about the mean at each time step. This interpretation can be directly visualized in the circled region in Figure 6, where the predicted values line up in a horizontal line. Although the model does not capture the time-dependent variations around the mean with high accuracy, its predictions are sufficiently reliable for estimating an individual's average VO_2max values during periods of peak power output.

The LSTM model's ability to detect patterns from time-series physiological signals makes it a favorable tool for approximating VO_2max-related metrics in a more accessible manner. Since the target value AO20 reflects oxygen uptake, the model's performance suggests that real-time fitness assessment could be feasible using wearable sensors alone in applied settings. This could be particularly impactful for athletes, clinicians, or fitness professionals seeking a practical, yet inexpensive and efficient method to assess aerobic capacity.

A study conducted by M. F. Akay et al. uses Generalized Regression Neural Networks (GRNN) to predict the VO_2max of an individual[25], which also showed high accuracy, achieving an R value of 0.81 and a standard error of estimate (SEE) of 4.87.

While direct comparisons are difficult, the higher R^2 values in this current study may be attributed to using an LSTM and its ability to capture long-term dependencies. It should be noted that the datasets and target variables used in the two studies are not identical; therefore, these results cannot be taken as proof of superiority, but rather as an indication that LSTMs can perform competitively in this research domain. At the same time, the high R^2 value reported in this study (0.9693) should be interpreted with caution. With only seven participants, the small dataset may bias the model's performance and limit its generalization ability, as highlighted within the red circle in Figure 6. In addition, the lack of an independent external dataset prevented validation beyond the original dataset. These results should therefore be interpreted as preliminary evidence supporting feasibility rather than conclusive evidence. Furthermore, no statistical significance testing was conducted due to the limited sample size. Statistical tests would have been underpowered and potentially misleading with only seven subjects. Future work with larger datasets would be able to support both external validation and statistical significance testing, helping establish generalizability.

However, many improvements can still be made to the model, such as improving the time-specific VO_2max predictions beyond the mean value. While the overall error remains low, improvements may be achieved by incorporating additional features, such as age or sex, to account for a more diverse sample of individuals. This demographic information is known to influence VO_2max. Future work could integrate it by concatenating a demographic feature vector with the LSTM's learned representation before the final feed-forward prediction. The present study deliberately excluded these features to emphasize a proof-of-concept scenario: testing whether VO_2max could be accurately predicted using only wearable physiological data without needing personal information or specialized lab equipment.

The data set consists of only seven individuals, which may introduce slight bias in the model and decrease its accuracy when generalizing the results to a larger, more diverse population. In addition to the small sample size, the dataset carries limitations that affect generalizability; namely, the demographic imbalance (six males and one female), the lack of diversity in age and fitness, the small sample size, and the use of AO20 as a surrogate for VO_2max. These limitations arise from the original structure of the dataset and should be considered when interpreting the results. Nonetheless, this research is intended as a proof-of-concept, demonstrating the feasibility and novelty of predicting VO_2max from wearable signals using an LSTM model.

Despite the limited dataset, the model demonstrated promising predictive performance without the need for enhancement techniques such as transfer learning. A randomized train/validation/test split was employed, and all reported results were obtained from the test set, following standard machine learning workflow. Expanding the dataset with more subjects and a wider range of physiological conditions could resolve this and enhance the model's ability to generalize the results to a larger population. These improvements can be made in future work, once more data is available.

Conclusion

This research aimed to develop a machine learning model that predicts VO_2max using physiological time-series data. Using an LSTM neural network, the project aimed to model temporal dependencies in exercise signals to approximate oxygen usage during recovery, helping with the analysis of aerobic fitness.

The study's methodology involved collecting incremental exercise data from seven subjects, creating labeled training data from 20-second rolling averages, and training the LSTM model on sequential inputs. After preprocessing and optimizing the architecture, the final model achieved an R^2 value of 0.9693, suggesting preliminary but favorable prediction ability in this proof-of-concept study. The model performed well due to the consistency of physiological responses during exercise recovery and the effectiveness of LSTM networks in modeling time-series data. The model is currently effective at predicting the mean VO_2max values during periods of peak power output.

Future work will focus on enhancing its ability to capture the time-dependent fluctuations around these mean values, expanding the dataset to encompass broader population variability. Given the limited size of the current dataset, simpler and more data-efficient approaches, such as ensemble methods or shallow neural networks, may also be explored as a baseline comparison to interpret the LSTM's results. Along with this, data augmentation techniques such as Gaussian noise injection, permutation of subsequences, magnitude warping, and synthetic oversampling methods like SMOTE for time-series may be employed to increase the functional sample size and reduce overfitting[26]. Demographic information could be incorporated into the model to improve generalization. Once more robust and diverse datasets become available, more complex architectures, such as transformers, could be evaluated. This model may then be implemented into an application or wearable device to help individuals understand their aerobic fitness without the need for lab testing, or be helpful to athletes whose training may be tailored based on their VO_2max.

References

1 K. Buttar, N. Saboo and S. Kacker, *A review: Maximal oxygen uptake (VO2 max) and its estimation methods.*

2 G. Rankovi, V. Mutavdi, D. Toski, A. Preljevi, M. Koci, G. Nedin-Rankovi and N. Damjanovi, *Aerobic capacity as an indicator in different kinds of sports.*

3 C. Gjestvang, T. Stensrud and L. Haakstad, *How is rating of perceived capacity related to VO2max and what is VO2max at onset of training?*

4 A. Smiley and J. Finkelstein, *Smart wearable analytics for cycling: AI-based physical exertion prediction*.

5 A. Smiley and J. Finkelstein, *Dynamic prediction of physical exertion: leveraging AI models and wearable sensor data during cycling exercise*.

6 S. Wiecha, P. Kasiak, P. Szwed, T. Kowalski, I. Cieliski, M. Postua and A. Klusiewicz, *VO2max prediction based on submaximal cardiorespiratory relationships and body composition in male runners and cyclists: a population study*.

7 C. Akalan, R. Robergs and L. Kravitz, *Prediction of VO2max from an individualized submaximal cycle ergometer protocol*.

8 A. Borror, M. Mazzoleni, J. Coppock, B. Jensen, W. Wood, B. Mann and C. Battaglini, *Predicting oxygen uptake responses during cycling at varied intensities using an artificial neural network*.

9 D. Svozil, V. b and J. Pospchal, *Introduction to multi-layer feed-forward neural networks*.

10 D. Poole and A. Jones, *Oxygen Uptake Kinetics*.

11 I. Mienye, T. Swart and G. Obaido, *Recurrent neural networks: a comprehensive review of architectures, variants, and applications*.

12 Z. Wang, Y. Song, S. Li and G. Sun, *Attention-enhanced CNN-LSTM model for exercise oxygen consumption prediction with multi-source temporal features*.

13 C. Hsiao, C. Tong and G. Cot, *Machine learning-based VO2 estimation using a wearable multiwavelength photoplethysmography device*.

14 R. Robergs, D. Dwyer and T. Astorino, *Recommendations for Improved Data Processing from Expired Gas Analysis Indirect Calorimetry*.

15 G. Houdt, C. Mosquera and G. Npoles, *A review on the long short-term memory model*.

16 A. Graves, *Long short-term memory. Supervised sequence labelling with recurrent neural networks*.

17 A. Zignoli, *Cycling VO2*, `https://www.kaggle.com/datasets/andreazignoli/cycling-vo2/data`.

18 *WMA declaration of helsinki - ethical principles for medical research involving human subjects*, World Medical Association.

19 A. Zignoli, A. Fornasiero, M. Ragni, B. Pellegrini, F. Schena, F. Biral and P. Laursen, *Estimating an individuals oxygen uptake during cycling exercise with a recurrent neural network trained from easy-to-obtain inputs: A pilot study*.

20 H. Jaafar, M. Rouis, L. Coudrat, E. Attiogb, H. Vandewalle and T. Driss, *Effects of load on wingate test performances and reliability*.

21 A. Castaeda-Babarro, *The wingate anaerobic test, a narrative review of the protocol variables that affect the results obtained*.

22 S. Mao and E. Sejdic, *A review of recurrent neural network-based methods in computational physiology*.

23 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, . Kaiser and I. Polosukhin, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 60006010,.

24 D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980.

25 M. Akay, E. etin, . Yarm, . Bozkurt and M. zilolu, *Development of novel maximal oxygen uptake prediction models for Turkish college students using machine learning and exercise data*, IEEE, p. 186189.

26 I. Annaki, M. Rahmoune and M. Bourhaleb, *Overview of data augmentation techniques in time series analysis*.