Oxygen Uptake Estimation During Cardiopulmonary Exercise Testing Using Temporal Fusion Networks

LUYAO YANG, CEMSE Division, King Abdullah University of Science and Technology, KSA
OSAMA AMIN, CEMSE Division, King Abdullah University of Science and Technology, KSA
AZMY FAISAL*, Department of Sport and Exercise Sciences, Manchester Metropolitan Institute of Sport, Manchester Metropolitan University, Manchester, M1 7EL, UK; Faculty of Sport Sciences for Men, Alexandria University, Egypt
BASEM SHIHADA*, CEMSE Division, King Abdullah University of Science and Technology, KSA

Accurate measurement of oxygen uptake ($\dot{V}O_2$) dynamics and maximal oxygen consumption ($\dot{V}O_2$ max), a vital marker of cardiorespiratory fitness and exercise capacity, requires specialized exercise physiology laboratories with costly equipment. This study develops a Temporal Fusion Network (TFN) approach utilizing easily accessible physiological parameters (heart rate, heart rate reserve, tidal volume, and breathing frequency), which can be measured with wearable sensors, anthropometric variables (age, gender, height, and weight), as well as health status to estimate $\dot{V}O_2$ dynamics during cardiopulmonary exercise testing (CPET). These input physiological parameters were derived from 140 laboratory CPET of a diverse cohort of adults (90 males, 50 females; 77 healthy, 63 smokers; average age: 26.6 years), to analyze $\dot{V}O_2$ dynamics. The TFN model demonstrated high predictive accuracy to estimate $\dot{V}O_2$ dynamics, with a Root Mean Square Error (RMSE) of 0.03 L/min and an R-squared (R2) value of 0.92, indicating robust performance across varied population groups. This TFN model paves the way for practical and cost-effective approach to estimate $\dot{V}O_2$ during exercise, with potential integration with consumer health devices to expand accessibility and, enhance its utility for clinical and fitness applications.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Oxygen Uptake, Deep learning, $\dot{V}O_2$ estimation, Temporal Fusion Network (TFN).

ACM Reference Format:
Luyao Yang, Osama Amin, Azmy Faisal, and Basem Shihada. 2024. Oxygen Uptake Estimation During Cardiopulmonary Exercise Testing Using Temporal Fusion Networks. *ACM Trans. Comput. Healthcare* 37, 4, Article 111 (May 2024), 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Oxygen uptake ($\dot{V}O_2$) dynamics during cardiopulmonary exercise testing (CPET) is the standard metric to evaluate the efficiency of the cardiovascular, respiratory and skeletal muscle systems to transport, and utilize

[*] Azmy Faisal and Basem Shihada are joint last authors.

Authors' Contact Information: Luyao Yang, luyao.yang@kaust.edu.sa, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA; Osama Amin, osama.amin@kaust.edu.sa, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA; Azmy Faisal, azmy.faisal@mmu.ac.uk, Department of Sport and Exercise Sciences, Manchester Metropolitan Institute of Sport, Manchester Metropolitan University, Manchester, M1 7EL, UK; Faculty of Sport Sciences for Men, Alexandria University, Egypt; Basem Shihada, Basem.shihada@kaust.edu.sa, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA.

ACM 2637-8051/2024/5-ART111
https://doi.org/XXXXXXX.XXXXXXX

oxygen in response to varying exercise intensities [1]. Maximal oxygen consumption ($\dot{V}O_2$ max) is derived from instantaneous $\dot{V}O_2$ measurements and represents the maximum oxygen uptake an individual can achieve during incremental CPET [1]. It is a quantitative benchmark of an athlete's or patient's aerobic fitness and is used to develop tailored training or rehabilitation plans [2]. Instantaneous $\dot{V}O_2$ measurements are used to estimate energy expenditure (EE), enhancing nutrition management for elderly patients frequently admitted to intensive care units [3, 4]. It also serves as a biomarker in medical research, allowing scientists to monitor the progression and status of diseases, including chronic cardiovascular and respiratory conditions [5, 6].

The common standard for measuring $\dot{V}O_2$ is cardiopulmonary exercise testing (CPET), which requires subjects to wear a mask or mouthpiece connected to a metabolic cart [7]. This equipment measures the volume and gas concentrations of inhaled and exhaled air. However, the bulky nature of the machinery and the need for professional operation in a lab setting limit its accessibility. Portable devices developed by companies like COSMED (Rome, Italy) [8] and VacuMed (USA) [9] offer more convenient testing options, but their high cost and the discomfort of wearing a facial mask still present barriers to widespread use, especially for prolonged periods during routine daily activities [10]. To overcome these limitations, some researchers have attempted to predict $\dot{V}O_2$ max using multivariate equations based on individual characteristics such as age, sex, weight, and height [11–14]. Additionally, commercial smartwatche products can estimate $\dot{V}O_2$ max [15], However, these methods often yield a single $\dot{V}O_2$ value and their models lack generalization across diverse populations. The significant error rates associated with these methods also limit their practicality in real-world settings [12, 15].

Recent advancements in wearable technology and artificial intelligence have provided new avenues to address health and exercise-related challenges [16]. Smart wearables have been validated for providing accurate measurements of physiological indicators such as heart rate (HR), electrocardiogram (ECG), and oxygen saturation (SpO_2) [17, 18]. Smart garments, for instance, can provide real-time measurements of respiratory parameters such as minute ventilation (VE) and breathing frequency (BF) [19]. These accessible parameters have been used in machine learning-based $\dot{V}O_2$ estimation and prediction models [20–23]. However, these methods exhibit several limitations. Primarily, they have relied on continuous physiological variables, such as VE, VT, and HR, while often omitting anthropometric variables like gender, height, and weight, which can enhance estimation models of $\dot{V}O_2$. Additionally, they tend to overlook the temporal dynamics inherent in time-varying variables, such as workload patterns that vary with different exercise protocols. Another critical limitation is the size of the datasets used for training and testing these models. For instance, the study conducted by Amelard et al., [21] used one of the largest datasets to estimate $\dot{V}O_2$, with a size of 22 participants, which may not adequately represent the broader population or the diversity of physiological responses. Furthermore, many models lack interpretability, failing to provide clear explanations for the impact of each input variable on the $\dot{V}O_2$ outcome.

To overcome the limitations of existing approaches, we employ a comprehensive temporal fusion model that leverages Long Short-Term Memory (LSTM) or Temporal Convolutional Network (TCN) as the encoder and employs attention mechanisms to ascertain the importance of input factors. Our model combines physiological parameters derived from laboratory CPET, that can be measured using wearable-sensors, with essential anthropometric information to estimate instantaneous $\dot{V}O_2$ throughout CPET with greater precision. Our key contributions are summarized as follows:

- We analyzed dataset of 140 participants which is the largest in literature to the best of our knowledge.
- We incorporate anthropometric variables such as age, gender, height, weight, in addition to health status, and physiological parameters VT, BF, HR, and HRR to enhance our model and its performance in estimating $\dot{V}O_2$ over various time spans.
- We design a TFN model leveraging physiological data that can be measured from cost-effective parameters measured wearable sensors in a non-invasive and continuous manner to estimate instantaneous $\dot{V}O_2$ during exercise and daily life activities.

- We account for temporal dynamics by including workload time-dependent variable workload data for each participant.
- We enhance the model's transparency by providing interpretable weights for each input variable, offering valuable insights into the relative importance of each input feature for $\dot{V}O_2$ estimation.

2 RELATED WORKS

The corpus of related literature encompasses numerous studies that have focused on estimating the singular $\dot{V}O_2$ max value by employing a range of statistical and machine learning methods [11–14, 24–28]. However, only a few delved into the realm of instantaneous $\dot{V}O_2$ estimation [21–23].

2.1 $\dot{V}O_2$ max estimation

Some research utilized questionnaires to gather basic information about individuals and then applied mathematical modeling or machine learning techniques to construct models. These models often incorporated anthropometric variables such as age, height, weight, gender, and maximal heart rate (HR max) to calculate $\dot{V}O_2$ max [11–14, 25]. Table 1 provides a comprehensive overview of these studies, detailing their methodologies and the input variables they used.

Table 1. Summary of methods for $\dot{V}O_2$ max estimation

Study	Methods	Input Variables
Frade et al. 2023 [25]	SVR	Sex, age, weight, height, and body mass index, breathing rate, minute ventilation, total hip acceleration, walking cadence, heart rate, and tidal volume
Petelczyc et al. 2023 [11]	Differential model	Gender, age, HR, HRmax, workload
Abut et al. 2019 [13]	SVM, GRNN, SDT	Gender, age, height, weight, HRmax, time, HR
Przednowek et al. 2018 [12]	SVM, MLP	Gender, distance, HRmax, recovery HR, age, weight, height, waist, hip, waist to height ratio, waist to hip ratio, BMI, fat mass index, fat-free mass index, body adiposity index, body surface area, fat, fat-free percentage, and total body water.
Abut et al. 2016 [14]	SVM, MLP	Gender, age, MX-HR, SM-ES, Q-PFA

2.2 Instantaneous $\dot{V}O_2$ estimation

Estimating instantaneous $\dot{V}O_2$ poses greater challenges than estimating $\dot{V}O_2$ max, because it involves capturing and understanding the complex dependencies and patterns present in $\dot{V}O_2$ sequential data. Determining $\dot{V}O_2$ max typically involves measuring peak oxygen uptake during a controlled test, while instantaneous $\dot{V}O_2$ requires continuous monitoring throughout an exercise session. This demands the resolution of notable technical challenges, and the variability of $\dot{V}O_2$ across different exercises and individuals complicates models. Here, we review recent methodologies aimed at estimating instantaneous $\dot{V}O_2$ and identify limitations within existing approaches, thus framing the context for our proposed solution.

One of the earliest attempts to quantify dynamic $\dot{V}O_2$ was by Su et al. in 2007, who employed support vector regression (SVR) to develop a model based on the pseudo-random binary sequence (PRBS) signal during running [29]. However, this approach was limited by its reliance on treadmill speed as the sole input, overlooking the inter-individual variations in physiological response. With the advancement of wearable technology, capable of capturing detailed physiological parameters, there has been an increase in studies utilizing machine learning and statistical methods to analyze data from these devices [20–23, 30].

Table 2. Summary of methods for estimating Instantaneous $\dot{V}O_2$

Study	Data size	Training data	Testing data	Methods	Model inputs	Device	Protocols or exercise
Su et al. 2007 [29]	6	6	6	SVR	PRBS and speed	Treadmill	Use PRBS to control the treadmill protocol
Altini et al. 2015 [22]	22	21	1(LOOCV)	SVM	ACC, HR, anthropometric features	ECG necklace, ACC	Lying, sedentary, dynamic, walking, biking, running
Cook et al. 2018 [30]	42	28	14	IAA	ECG, ACC, HR	DREEM, COSMED K4b^2	Bruce protocol
Zignoli et al. 2020[20]	7	7	7	LSTM	HR, RF, P, ω	power meter, COSMED	Arbitrary protocols, Wingate test
Shandhi et al. 2020 [23]	17	16	1(LOOCV)	XGBoost	SCG, ECG, AP	Custom-built wearable patch	Treadmill protocol, outside protocol
Amelard et al. 2021 [21]	22	17	5	TCN	HR, HRR, RF, VE	Smart shirt	One ramp-incremental, PRBS protocol
Our study	140	100	40	TFN	Gender, age, height, weight, workload, HR, HRR, VT	Vyntus CPX, Cortex Metalyzer	Incremental exercise

Altini et al. [22] and Shandhi et al. [23] employed the Leave-One-Out Validation (LOOV) technique to evaluate their models. On the other hand, Zignoli et al. [20] trained their model using two protocols per individual and then tested it on a distinct protocol for each person.

Altini et al. were among the first to use accelerometer (ACC) and HR sensor data to estimate instantaneous $\dot{V}O_2$ during various daily activities, including lying, sitting, walking, biking, and running. By leveraging support vector machine (SVM), they developed a range of models specifically tailored to estimate instantaneous $\dot{V}O_2$ during different activities [22]. However, the need to create models for each activity posed a challenge in terms of universality and practical application. Then, Cook et al. designed a mathematical algorithm combining HR with the integral of absolute acceleration (IAA) to estimate instantaneous $\dot{V}O_2$ [30]. The reliance on a mathematical framework may introduce biases and limit the model's ability to reflect the complexity of real-world data.

To make the model more applicable in different exercises, Shandhi et al. utilized a custom-built wearable patch placed on the mid-sternum to collect seismocardiography (SCG), electrocardiogram (ECG), and atmospheric pressure (AP) signals from 17 adults using a treadmill protocol within a controlled setting, as well as an outdoor walking protocol in an uncontrolled environment [23]. Later, they trained the eXtreme Gradient Boosting (XGBoost) models on one protocol of each person and validated the data from the other protocol and vice versa.

Moreover, neural networks demonstrate a robust ability to learn features directly from raw data. As a result, there are also studies employing Long Short-Term Memory (LSTM) networks to estimate $\dot{V}O_2$ [20]. This study collected the HR, breathing frequency (BF), mechanical power output (P), and pedaling cadence (ω) of 7 amateur cyclists in 3 protocols (two arbitrary protocols and the Wingate test). However, this study relied on a power meter rather than a wearable sensor, restricting its applicability to cycling. During the process, Two protocols of each person are used as the training set, and the remaining protocol is used as the test set.

Amelard et al. conducted the first research using temporal convolutional network (TCN) to predict $\dot{V}O_2$ based on the smartshirt data, showcasing the capabilities of deep learning in predicting instantaneous $\dot{V}O_2$ [21]. They first collected smart shirt data (HR, HR reserve, BF, and minute ventilation (VE)) from 22 adults. Based on the temporal behavior of the data, they tried to train a TCN on 17 adults and test the model on the rest 5 adults [21]. This work shows the great power of deep learning in prediction the instantaneous $\dot{V}O_2$.

In summary, while existing methods have significantly advanced the instantaneous $\dot{V}O_2$ estimation, it is important to acknowledge that they have some limitations. Firstly, they ignored the physical and anthropometric features of each individuals, which are critical determinants of $\dot{V}O_2$. Furthermore, although Frade et al. [25] employed SHAP to elucidate the input variables of their $\dot{V}O_2$ max prediction model, there is a notable scarcity of research that similarly addresses the interpretability of input variables in the prediction of instantaneous $\dot{V}O_2$ Moreover, the reliance on small datasets tends to assume a lack of diversity within the population studied.

To address these issues, as detailed in Table II, we have compiled a more extensive dataset comprising 140 participants, with 100 allocated for training and 40 for testing. Our objective is to integrate both anthropometric (such as gender, health status, age, height, weight) and temporal features (such as workLoad, HR, HRR, BF, VT) collected from wearable devices into a deep learning model capable of making accurate $\dot{V}O_2$ estimation. We have validated our model across diverse groups, including young and elderly participants as well as smokers and non-smokers, to ensure that it performs well and possesses robust generalization capabilities.

2.3 Multi-horizon Estimation Models

In multi-horizon estimation tasks, the challenges often arise from the complex amalgamation of various inputs. These inputs comprise static covariates, including attributes like height and weight, which remain constant throughout the time series. Additionally, future inputs such as workLoad are known in advance, while exogenous time series data, such as HR, BF, and VT, are solely available in historical records. However, in these scenarios, a common issue is the absence of prior knowledge about the relationships and interactions among these inputs and the target variable such as $\dot{V}O_2$.

To solve this, various deep learning methods have emerged to address this task, including methods such as autoregressive models and sequence-to-sequence models. Autoregressive methods have been widely used in this field [31–33]. They capture the dependencies within a time series by regressing the current value on its past values. Some methods such as DeepAR [31] and Deep State-Space Models (DSSM) [32] employ LSTM networks to capture temporal patterns and dependencies, enabling them to generate probabilistic estimate with measures of uncertainty. Some transformer-based methods used the Convolutiuonal Neural Network (CNN) as the local processing for the forecasting [33]. They offer the advantage of capturing temporal patterns and trends in the data. However, autoregressive methods may necessitate either stable (stationary) data or data that becomes stable through differencing.

Sequence-to-sequence models, on the other hand, differ in that they are trained to explicitly produce forecasts for multiple predetermined horizons at each time step, along with various techniques to generate future predictions. The Multi-horizon Quantile Recurrent Forecaster (MQRNN) [34] adopts LSTM or CNN encoders to generate context vectors that are subsequently fed into Multi-Layer Perceptrons (MLPs) specific to each forecasting horizon. The Temporal Fusion Transformers (TFT) [35] combines the strengths of Transformers and LSTM networks to capture temporal patterns and dependencies in time series data.

3 METHODS

$\dot{V}O_2$ kinetics exhibit complex temporal patterns in response to exercises, making it challenging to capture the temporal information. To address this, we have developed a novel model called Temporal Fusion Network (TFN), which aims to enhance interpretability and accuracy. Our TFN model consists of three main components: feature selection, temporal extraction, and output generation. Inspired by [35], we built a feature embedding module that combines and fuses the anthropometric and dynamic input information into the input embeddings. During the training process, we dedicate an initial stage to continually updating the ratio of each parameter to elucidate the relative importance of different variables within the model. Subsequently, we establish a temporal extraction module that learns the temporal features of the input sequences. Lastly, the output generation component will be learned to estimate the $\dot{V}O_2$. In the following subsections, we present a detailed understanding of the method implementation and model structure.

3.1 Data Collection and Preprocessing

3.1.1 Subjects. This study involved 140 participants (90 males, 50 females; 77 healthy, 63 smokers; height: 175.9 cm ± 2.9 cm weight; 75.7 kg ± 3.6 kg; age: 26.6 ± 2.7 years, 88% are Caucasians). Prior to participating in the study, all subjects provided written informed consent and confirmed they had no cardiopulmonary conditions.

3.1.2 Data Collection. Metabolic measurements were obtained using the Vyntus CPX system (Vyaire Medical, Hochberg, Germany) and the Cortex Metalyzer 3B system (Cortex, Leipzig, Germany) at Manchester Metropolitan Institute of Sport, Manchester Metropolitan University, UK. These measurements included key parameters such as $\dot{V}O_2$, VT, and BF. Heart rate was continuously monitored throughout the assessment using the Polar H7 tooth heart rate monitor (Polar, Kempele, Finland), ensuring accurate tracking of cardiovascular responses during exercise. The assessments were conducted on a cycle ergometer, facilitating a controlled evaluation of the participants' metabolic responses during incremental exercise.

Prior to testing, participants underwent a thorough medical screening to confirm their suitability for high-intensity exercise. They were instructed to adhere to a pre-test protocol, which required them to refrain from engaging in strenuous physical activity, consuming caffeine, and eating large meals for at least three hours before the assessment.

The testing protocol began with a five-minute rest period to effectively prepare participants physically and mentally for the upcoming challenge. After the rest period, the main testing phase began, characterized by a gradual increase in workload of 20 watts every two minutes, starting with an initial 0 watt workload. This progressive approach continued until each participant reached their $\dot{V}O_2$ max, providing a comprehensive understanding of their metabolic function and exercise capacity.

After the test is completed, post-test procedures include a carefully monitored two-minute cool-down period, which is critical to ensure the safe recovery of the participant. Subsequent data analysis is conducted to accurately determine the instantaneous $\dot{V}O_2$ values as well as $\dot{V}O_2$ max, which was determined as the average $\dot{V}O_2$ values recorded during the last 15s at peak exercise work rate. $\dot{V}O_2$ max was attained by the participants achieving the following criteria for maximal effort: A RER > 1.10, and HR >90% HRmax predicted for age [36]. Participants maintained a constant cadence of 60 rpm and the test was terminated when the subject was unable to continue or maintain > 60 rpm cadence. Throughout the testing process, stringent safety protocols are upheld to ensure any adverse reactions are addressed immediately, maintaining the highest standards of participant safety and test integrity.

3.1.3 Data Preprocessing.

In the initial stages of raw data preprocess, we compute the mean μ^i and standard deviation σ^i of the variables for i-th participant across the temporal dimension. Subsequently, we remove the data points that fall below the threshold of $\mu^i - 2 * \sigma^i$ and above the threshold of $\mu^i + 2 * \sigma^i$. Upon cleaning the data, these significant outliers were excised from the dataset to ensure the integrity of further data processing steps. Following the removal of outliers, we implemented a data sampling strategy to systematically select representative data points. This was achieved by resampling each point at regular two-second intervals, utilizing linear interpolation to estimate missing values where necessary.

While the aforementioned steps have addressed the issue of obvious outliers, there remained the possibility of less apparent anomalies or noise within the data. To tackle this, we employed a data smoothing technique rolling window. Specifically, a window size of two was chosen to average the data points within each window, effectively reducing short-term fluctuations and highlighting longer-term trends or cycles. Suppose we have a time series $x = x_1, x_2, ..., x_n$ and we want to apply a rolling window of size k. Then the smoothed value y_i at time i can be calculated as:

$$y_i = \frac{1}{k} \sum_{j=0}^{k-1} x_{i-j} \tag{1}$$

Unlike previous methodologies that collect data within a fixed temporal span [21, 23], thereby enforcing uniform exercise and rest durations for all subjects, our data collection approach embraces individual variability in time lengths. An illustrative example of this variability can be observed in the gender-based discrepancy where men often require more time than women to attain their $\dot{V}O_2$ max.

The average test duration per participant in our dataset is 762 seconds. To enhance the dataset and standardize the input size for our model, we employ a sliding window approach. In our specific experiments, we denote the sequence length of the input variable for an individual as S^i, where i represents the unique identifier of that individual. We utilize a sliding window of length $e = 200$ seconds with a moving step of half the window size, specifically 100 seconds. This configuration allows us to generate a substantial number of windowed datasets. As a result, we obtain $S^i - e + \frac{e}{2}$ time span samples for each individual i. This method not only expands our dataset but also effectively addresses the inherent variability in sequence lengths across the dataset.

To ensure a robust evaluation of our model, we employed a stratified sampling method based on different age groups, dividing our dataset into homogeneous subgroups, each representing a decade of age. Specifically, we divided our dataset into homogeneous subgroups based on different age groups (one age group for every 10 years old). We then split each subgroup into a training dataset and a test dataset, maintaining a ratio of 0.7 to 0.3 respectively. This stratified approach ensures that our training and testing sets accurately and comprehensively reflect the overall age distribution of our dataset, allowing us to better evaluate the performance of our model across different age groups.

Finally, we employ standard normalization to normalize the continuous variables, which include BF, HR, HRR, VE, VT. This normalization process effectively removes the mean and scales the data to unit variance, thereby facilitating more efficient model training and enhancing convergence speed. Concurrently, we utilize ordinal encoding [37] to convert categorical variables, such as ID, status, height, weight, gender, and age, into integer representations. This preprocessing step is crucial for improving both the performance and interpretability of the model in subsequent analyses, ensuring that all input features are adequately prepared for integration into the machine learning framework.

3.2 Sequence Modeling

Before defining the network structure, we highlight the nature of the multi-horizon forecasting sequence modeling task. Consider a set of time-series input sequences $X^i = x_1^i, x_2^i, \ldots, x_t^i, \ldots, x_T^i \in \mathbb{R}^d$, $i \in \{0, 1, 2, ..., n\}$. n represents the number of input variable types. It can either include both dynamic and anthropometric data (n=12) or only dynamic data (n=6). Each input sequence, denoted as X^i, is composed of T timesteps, with each timestep t having as associated d dimension vector x_t^i After feature selection, the combined embedded inputs formed by concatenating both the anthropometric variables and dynamic variables:

$$x_1, x_2, \ldots, x_T = \text{concat}\left(\bigcup_{i=0}^{n}\{x_1^i, x_2^i, \ldots, x_T^i\}\right) \tag{2}$$

After the concatenation, the model $\mathcal{M}$ aims to estimate the ground truth sequence of $\dot{\text{V}}\text{O}_2$ values denoted as $\dot{\text{V}}\text{O}_2 : \dot{\text{V}}\text{O}_{2(1)}, \dot{\text{V}}\text{O}_{2(2)}, \dot{\text{V}}\text{O}_{2(3)}, \ldots, \dot{\text{V}}\text{O}_{2(T)}$ values from the wearable input time-series $x_1, x_2, \ldots, x_T$, which is formalized as follows,

$$\widehat{\dot{\text{V}}\text{O}_2} = \mathcal{M}(x_1, x_2, x_3, \ldots, x_T) \tag{3}$$

The training process aims to minimize the divergence between the model's estimated $\widehat{\dot{\text{V}}\text{O}_2}$ outputs and the actual $\dot{\text{V}}\text{O}_2$ values.

3.3 Model Architecture

Our TFN model is based on the sequence-to-sequence model architecture, as shown in Fig. 1. In this section, we describe our architecture primary parts: feature selection and temporal extraction for encoding the input variables, and output generation for decoding the extracted information.

3.3.1 *Feature Selection.*

We utilize entity embeddings [37] to encapsulate static variables (Gender, Age, Status, Height, Weight), while we employ linear transformations to encode dynamic variables (HR, HRR, VT, BF, WorkLoad). Let c denote the transformed concatenated static variables, and let θ_t denote the transformed concatenated dynamic variables corresponding to the input x_t at time-step t. To adequately combine the temporal dependencies and dynamics within the input variables and select the important ones, we apply the variable selection module GRN proposed by lim et al. [35] to the transformed concatenated input variables, followed by a Softmax layer:

$$\gamma_t = \text{Softmax}(\text{GRN}(\theta_t, c)) \tag{4}$$

After incorporating the variable selection module, we acquire the input featuring weighted embeddings $\gamma_t \in \mathbb{R}^d$, where $\mathbb{R}$ represents the set of real numbers, t denotes the time-step, and γ_t is a d-dimensional vector that combines both static and dynamic information. This part helps to enhance model transparency while facilitating the identification of key determinants among the input variables.

3.3.2 *Temporal Extraction.*

In this part, temporal attributes are acquired through the implementation of sequence modeling techniques. Within the domain, multiple investigations have employed RNN or CNN architectures as encoders to capture the fundamental patterns in γ_t. In a manner akin to MQRNN [34], we employ two modules within this section. One module adopts LSTM as the encoder to extract temporal information, while the other module utilizes TCN as the encoder [38], serving the same purpose of capturing temporal characteristics. Both TCNs and LSTMs have proven to be effective in various time series analysis tasks, such as speech recognition, natural language processing, and sequential data forecasting. In our experiment, both structures show competitive results. We will compare the results of these two modules on the test dataset in Section IV.

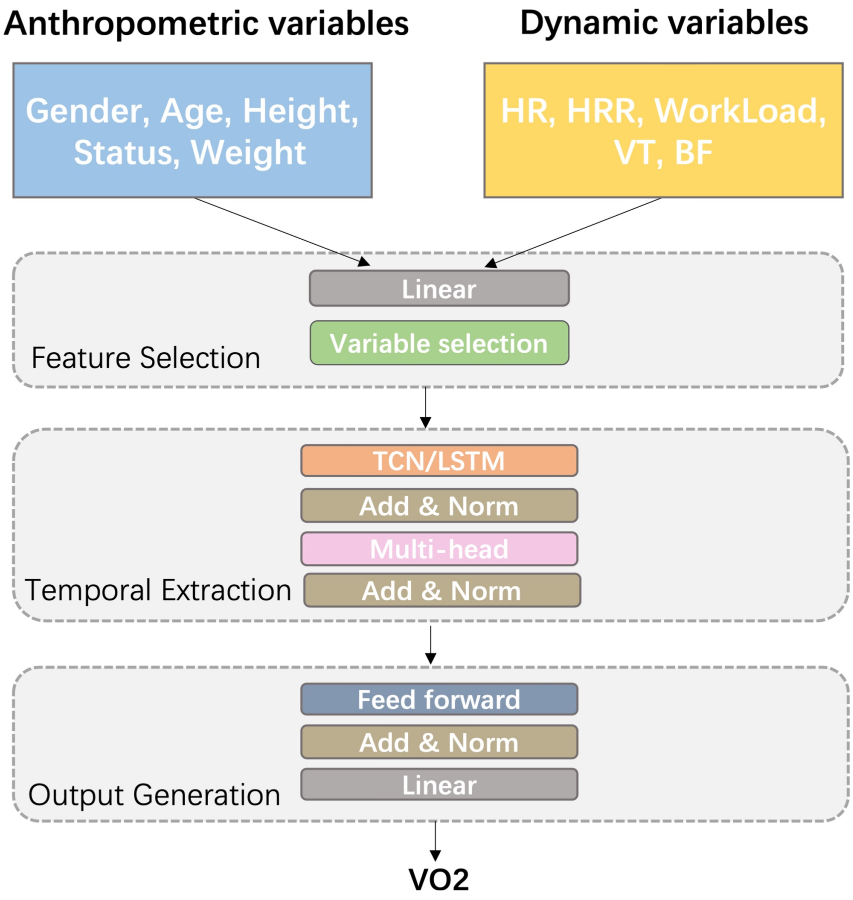

Fig. 1. Schematic representation of the TFN model. Add & Norm represents the skip connection and layer normalization, respectively. Multi-head represents the multi-head attention mechanism. The model consists of three main sections: Variable Selection uses two separate networks to process anthropometric and dynamic variables from the input data. Temporal Extraction uses a TCN or LSTM, and to understand patterns and importance over time. The final part, Output Generation, uses a feed-forward network to estimate the $\dot{V}O_2$ values.

We denote the TCN or LSTM model as Module. The Module receives γ_t and then extracts time-dependent information in its respective encoding module, ultimately producing the transformed vector ξ_t at time t:

$$\xi_t = \text{Module}(\gamma_t) \tag{5}$$

In order to enhance the efficiency of training, γ_t is connected to ξ_t via a residual skip connection, which is represented by $\oplus$. Subsequently, layer normalization is applied, yielding the refined temporal embedding denoted as $\tilde{\xi}_t$:

$$\tilde{\xi}_t = \text{LayerNorm}\left(\text{Module}(\gamma_t) \oplus \gamma_t\right), \tag{6}$$

which facilitates gradient flow and lead to an effective model training. Subsequent to the normalization step, we apply a multi-head self-attention layer, denoted as Atten to $\tilde{\xi}_t$ obtaining ψ_t as,

$$\psi_t = \text{Atten}(\tilde{\xi}_t). \tag{7}$$

The utilization of multi-head attention allows the model to effectively capture the relationships between different positions within the sequence $\tilde{\xi}_t$. In our setup, to capture a richer representation of the data, we employed two heads in the multi-head attention mechanism.

In a manner consistent with previous steps, we integrate a skip connection to merge the temporal embeddings ψ_t and $\tilde{\xi}_t$. Then, it is followed by a normalization process to align with the procedure outlined in (6) to get the extracted high-level feature representations $\tilde{\psi}_t$,

$$\tilde{\psi}_t = \text{LayerNorm}(\tilde{\xi}_t \oplus \text{Atten}(\tilde{\xi}_t)) \tag{8}$$

By employing the aforementioned approach, we ensure the stability and effectiveness of the model's training process. Regarding the implementation phase, the details of the model's parameters design are elaborated in Section IV.

3.3.3 Output Generation. After processing and transforming the input data through variable selection and temporal extraction parts, $\tilde{\psi}_t$ are passed to the final Feed-Forward Network (FFN) to get the transformed output ϕ_t. The FFN contains two fully-connected layers that act as a non-linear regressor, projecting the encoded multi-dimensional feature space onto the target space of $\dot{\text{V}}\text{O}_2$ values.

Following a similar approach as in the temporal extraction part, we obtain a transformed representation, $\tilde{\phi}_t$, by incorporating residual skip connections and layer normalization techniques. These techniques facilitate establishing connections between ϕ and $\tilde{\psi}_t$, enhancing the integration and stability of the data transformation process.

Finally, a linear layer is utilized to output the estimated $\widehat{\dot{\text{V}}\text{O}}_{2t}$ values:

$$\begin{aligned} \tilde{\phi}_t &= \text{LayerNorm}(\tilde{\psi}_t \oplus \text{FFN}(\tilde{\psi}_t)) \\ \widehat{\dot{\text{V}}\text{O}}_{2t} &= \text{Linear}(\tilde{\phi}_t) \end{aligned} \tag{9}$$

By fusing all available input information, the FFN performs the task of estimating $\dot{\text{V}}\text{O}_2$ using the linear layer. It ties together all the transformations performed by the preceding interpretable variable selection and representation learning stages to distill them into informative $\dot{\text{V}}\text{O}_2$ estimations.

3.4 Loss function

Quantile loss function is frequently employed in time series prediction scenarios where the objective is to estimate an interval, or quantile, rather than a single point estimate. Specifically, the quantile loss function is designed to estimate the τ-th quantile of the conditional distribution of the response variable, where τ represents the desired quantile level. By varying τ, the model can provide estimates with different levels of confidence, which can be useful when dealing with uncertainty and variability in time series data.

The quantile loss at i-th timestep is expressed as,

$$\begin{aligned} L_\tau(\dot{\text{V}}\text{O}_{2(i)}, \widehat{\dot{\text{V}}\text{O}}_{2(i)}) = \tau * \max(\dot{\text{V}}\text{O}_{2(i)} - \widehat{\dot{\text{V}}\text{O}}_{2(i)}, 0) + \\ (1-\tau) * \max(\widehat{\dot{\text{V}}\text{O}}_{2(i)} - \dot{\text{V}}\text{O}_{2(i)}, 0). \end{aligned} \tag{10}$$

To compute the total loss across all N timesteps, the individual quantile losses are summed. Here, N represents the total number of timesteps, $\tau_{\max}$ denotes the total number of quantiles, and the double sum spans all timesteps

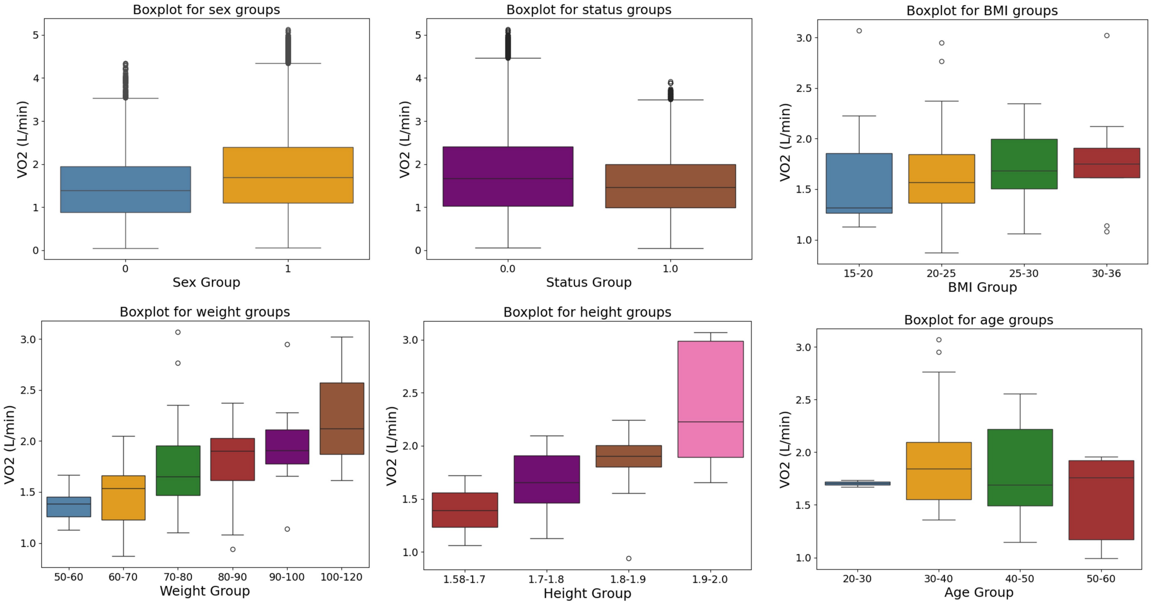

Fig. 2. For each specific anthropometric variable, the mean $\dot{V}O_2$ value was calculated within each group. These calculations revealed noticeable disparities between the groups, indicating the distinct influence of each variable on $\dot{V}O_2$ values.

I and quantiles τ, as demonstrated in Equation (11).

$$L_\tau(\dot{V}O_2, \widehat{\dot{V}O_2}) = \sum_{\tau=1}^{\tau_{\max}} \sum_{i=1}^{N} \frac{L_\tau(\dot{V}O_{2(i)}, \widehat{\dot{V}O_{2(i)}})}{N\tau_{\max}} \tag{11}$$

Quantile loss allows the model to assess estimates over a range of possible outcomes, which can provide more information than a simple point estimation assessment. We use three various percentiles (e.g., 10th, 50th, and 90th) at each timestep, which means the $\tau = 0.1, 0.5, 0.9$. Ultimately, we employ the output values corresponding to the 50th percentile ($\tau = 0.5$) as our final outputs, capturing the central tendency of the distribution while accounting for the uncertainty inherent in the estimation.

3.5 Evaluation Metrics

We employ the Root Mean Square Error (RMSE) and the R-squared (R^2) as evaluation metrics. Let $\dot{V}O_{2(i)}$ be the true value and $\widehat{\dot{V}O}_{2(i)}$ be the estimated value at the i-th timestep, the RMSE can be calculated from,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\dot{V}O_{2(i)} - \widehat{\dot{V}O}_{2(i)}} \tag{12}$$

Here, $n = 140$, which corresponds to the number of participants. The MAE is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\dot{V}O_{2(i)} - \widehat{\dot{V}O}_{2(i)}| \tag{13}$$

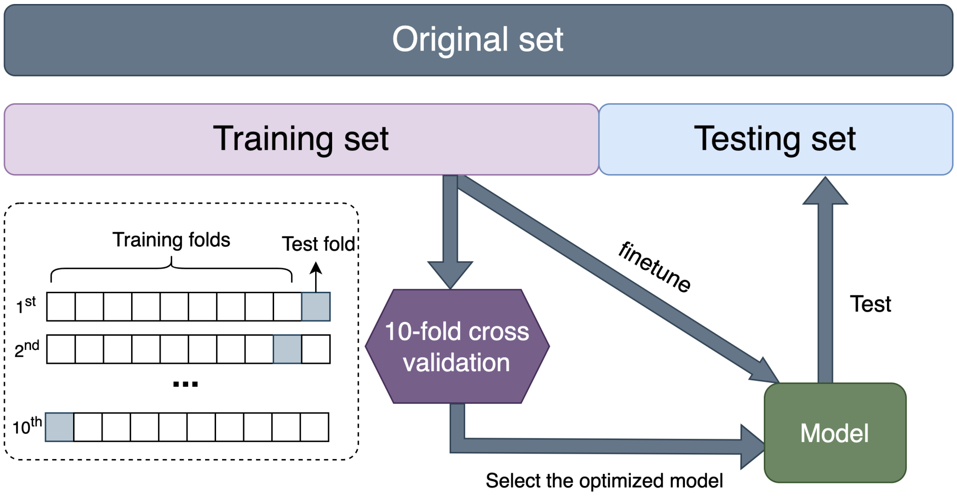

Fig. 3. The original dataset was divided into training and test sets with a ratio of 7:3. Subsequently, 10-fold cross-validation was conducted on the training set to identify the optimal model. Finally, the selected model was fine-tuned and validated using the test set.

R^2 provides a quantitative measure of how well the estimated $\widehat{\dot{V}O_2}$ values from a model align with the actual values $\dot{V}O_2$. It is defined as the proportion of the variance in the dependent variable (in this case, the true $\dot{V}O_2$) that is explained by the independent variable(s) (in this case, $\widehat{\dot{V}O_2}$). A higher R^2 indicates a better fit of the model and suggests that the model can better explain the variation in the data. Firstly, we calculate the mean of the true $\dot{V}O_2$ values, denoted as c:

$$\overline{\dot{V}O_2} = \frac{1}{n}\sum_{i=1}^{n}\dot{V}O_{2(i)} \tag{14}$$

Compute the total sum of squares (TSS), which is the sum of squares of the difference between the true $\dot{V}O_2$ value and $\overline{\dot{V}O_2}$, to quantify the total variance in the data as,

$$TSS = \sum_{i=1}^{n}(\dot{V}O_{2(i)} - \overline{\dot{V}O_2})^2. \tag{15}$$

Then, calculate the residual sum of squares (RSS), which is the sum of squares of the difference between the true $\dot{V}O_2$ value and the $\widehat{\dot{V}O_2}$ value. This measure quantifies the variance that the model fails to explain,

$$RSS = \sum_{i=1}^{n}(\dot{V}O_{2(i)} - \widehat{\dot{V}O}_{2(i)})^2 \tag{16}$$

Finally, based on the (15) and (16), we could calculate R^2 using the formula:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{17}$$

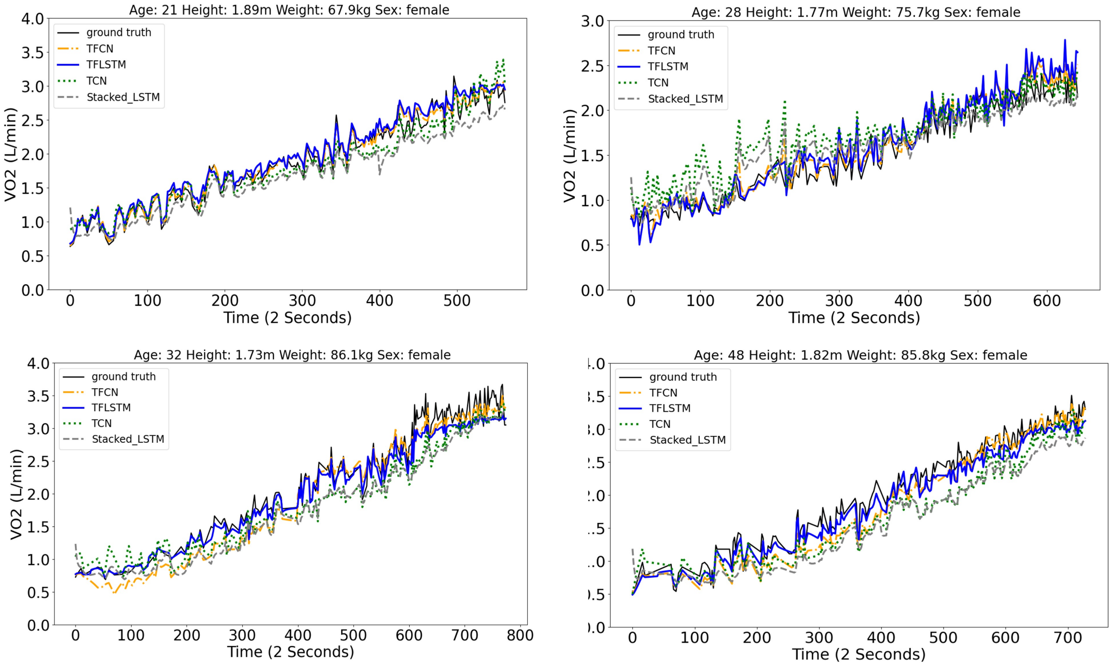

Fig. 4. We present four optimal scenarios across different age groups selected for comparison. The TFCN (orange) and TFLSTM () demonstrate superior performance in estimating the true $\dot{V}O_2$ values (black) compared to the TCN (green) and LSTM (gray) models. It is noteworthy that all four models were trained on the same datasets, which included both dynamic and anthropometric variables.

3.6 Validation Experiments

To further validate the effectiveness of our approach, we conducted two experiments, one of which employed 10-fold cross-validation, as illustrated in Fig. 3. Initially, the dataset was divided into training and test subsets in a 70:30 ratio, with 70% for training and 30% for testing. This division followed the age stratification of the participants to ensure representative subsets. Subsequently, we performed 10-fold cross-validation on the training set, partitioning it into 10 folds. For each fold, the model was trained on 9 folds while being validated on the remaining fold. Throughout this process, we assessed various model configurations and hyperparameters to identify the best-performing model. Upon determining the optimal model and its associated hyperparameters, we fine-tuned the model using the entire training set. Finally, we evaluated the model's performance on the separate test set, providing an unbiased estimate of its effectiveness on unseen data and ensuring the robustness of our results.

Additionally, we conducted two separate ablation studies. The first study investigated the impact of introducing various levels of noise to the original dataset on the model's predictive performance. The second study examined the influence of incorporating anthropometric data as input variables on the model's overall effectiveness. Together, these studies provide valuable insights into how both noise and anthropometric factors affect model performance.

3.7 Model Optimization

The optimization process employs the Adam optimizer, selected for its efficiency in managing sparse gradients and its adaptive learning rate capabilities, which facilitate effective training across various data distributions. Both the TCN and LSTM modules are configured with two layers, enabling the model to capture complex temporal dependencies inherent in the data. The training regimen spans 50 epochs, utilizing a learning rate of 0.01 and a batch size of 64, which together enhance the model's convergence while maintaining computational efficiency. To mitigate the risk of overfitting, a dropout rate of 0.1 is incorporated as a regularization technique. Additionally, an embedding dimension of 16 is utilized to represent categorical features in a continuous space, and the inclusion of two attention heads further enriches the feature representation, allowing the model to focus on relevant information within the input sequences. This comprehensive configuration is designed to optimize the model's predictive performance in time-series forecasting tasks.

4 RESULTS

4.1 Static Variables Distribution

The distribution of static variables in the entire dataset is presented in Fig. 2, which illustrates the mean distribution of the $\dot{V}O_2$ variable with respect to various anthropometric parameters, including gender, age, status, height, weight, and BMI. The figure provides a visual representation of these anthropometric data characteristics, facilitating a more comprehensive understanding of the dataset.

The proposition that $\dot{V}O_2$ is influenced by age, attributed to the decrease in metabolically active tissue associated with aging, has been in consideration since 1988 [39]. Concurrently, sports scientists began to recognize the effect of factors such as gender, weight, and BMI on $\dot{V}O_2$ [40]. This is intuitively comprehensible, as a larger body necessitates a greater oxygen supply for its functioning. Moreover, empirical observations have revealed that, on average, males tend to exhibit higher $\dot{V}O_2$ ranges compared to females. Furthermore, it has been observed that individuals who are in good health tend to demonstrate higher $\dot{V}O_2$ values in comparison to smokers. Due to the significant relationship between static variables and $\dot{V}O_2$, they were utilized into early methodologies for constructing estimations of $\dot{V}O_2$.

Therefore, our proposed model leverages static variables as informative priors for the model, which are then integrated with dynamic variables. This integration enables the model to enhance its capabilities in estimating the $\dot{V}O_2$ levels of diverse individuals.

Table 3. Comparison of RMSE(L/min), MAE(L/min), and R^2 for different methods

Methods	R^2 ↑	RMSE ↓	MAE ↓
TCN	0.81	0.12	0.23
Stacked LSTM	0.82	0.10	0.22
Base TFCN	0.85	0.07	0.19
Base TFLSTM	0.86	0.06	0.19
TFCN	0.91	0.03	0.14
TFLSTM	**0.92**	**0.03**	**0.13**

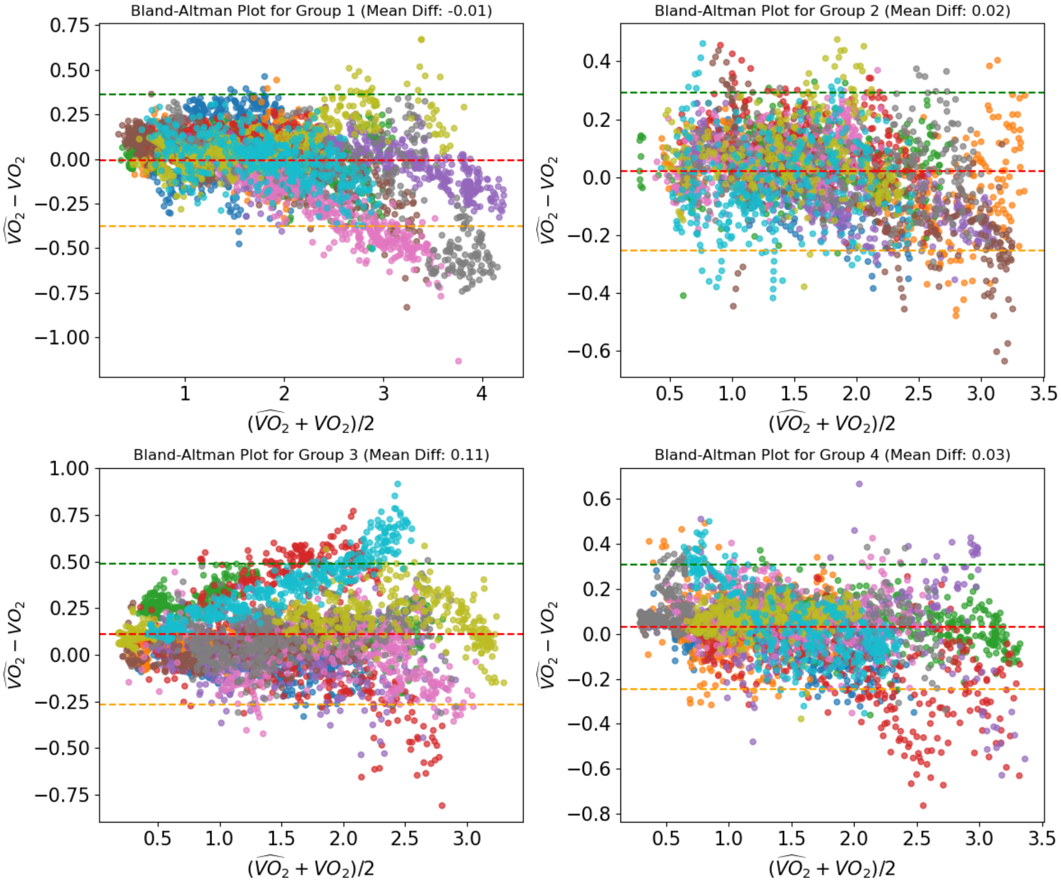

Fig. 5. The four figures illustrate the Bland-Altman analysis for the four age groups. The dotted horizontal line indicates the 95% consistency limit, while the red dotted line denotes the prediction deviation. The deviation values for each of the four figures are included in the titles. Different colors in the figures represent data from various participants within the test set.

4.2 Comparison with other methods

In prior research, various studies have explored the utilization of LSTM and TCN for the estimation of $\dot{V}O_2$ [20, 21]. However, a conspicuous gap in the literature is the lack of available code for these models, thereby limiting comparative analysis of their performance. In this study, we implemented the LSTM and TCN model used in [20, 21] for comparing the performances of different methods. For our TFN model, we used two distinct models were used for this purpose:

- the "Base model": the model designed to use only dynamic variables inputs.
- the "model": the model formulated to use both dynamic and anthropometric variables as inputs.

Additionally, we employ TCN and LSTM as the fundamental modules for the Temporal Extraction part of our model. Consequently, we construct two distinct models, named TFCN and TFLSTM, by incorporating TCN and LSTM, respectively, in the Temporal Extraction part.

Table 4. Variables Importance Rank

(a) Variables Importance in TFCN

Anthropometric Variable	Importance	Dynamic Variable	Importance
Height	0.24	VT	0.28
Weight	0.24	WorkLoad	0.21
Status	0.19	BF	0.19
Sex	0.17	HRR	0.17
Age	0.16	HR	0.15

(b) Variables Importance in TFLSTM

Anthropometric Variable	Importance	Dynamic Variable	Importance
Age	0.22	VT	0.27
Height	0.21	BF	0.24
Sex	0.21	WorkLoad	0.21
Weight	0.19	HRR	0.15
Status	0.17	HR	0.13

As shown in Fig. 3, we initially performed 10-fold cross-validation [41] on 70% of the training set, resulting in an average model performance of R^2: 0.95 ± 0.01, RMSE: 0.18 ± 0.01, and MAE:0.13 ± 0.01. Based on these results, we identified the best-performing model and subsequently retrained it using the entire training dataset. We then compared the performance of our models with other methods on an independent test set. The performance of all models is evaluated based on MAE, RMSE, and R^2 values, as shown in Table 3. Among the various methods tested, our TFN model, particularly the TFLSTM model, exhibited state-of-the-art performance. We evaluated all the models on the 40 testing files. Fig. 4 shows the good scenarios in different age groups, TFN shows the robustness compared with other models. In addition, we performed a Bland-Altman analysis on the test set. As shown in Fig. 5, the results show that our model achieves more than 95% consistency in the four age groups in the test set.

In Fig. 4, it can be observed that TCN and Stacked LSTM models exhibit higher susceptibility to noise, resulting in some erroneous estimations (e.g. in the 100-200 time period of Example 1). Conversely, our proposed models, TFLSTM and TCN, demonstrate enhanced robustness in the presence of noise, leading to more accurate performance. Our model improved resilience to noise-induced fluctuations, thereby yielding more reliable and precise results.

4.3 Variables Interpretation

Apart from superior accuracy, our model offers explanatory insights into anthropometric and dynamic variables, the weights are obtained by part Feature Selection. By averaging each variable's importance through all participants, we get the importance value assigned to each input variable serves as a measure of its significance to the final $\dot{V}O_2$ values estimations. Table 4a and Table 4b present the respective variable importance rankings for the TFCN and TFLSTM models. These tables highlight the significance and contribution of different variables within each model, providing valuable insights into the relevance of the variables in the context of the models' performance.

Among the five static variables analyzed, their importance scores consistently hover around 0.2, indicating that each variable contributes equally to the model's predictive capacity. This finding further corroborates the insights presented in Fig. 2, which illustrates that the various static variables possess significant prior knowledge regarding the dataset. In both models, the dynamic variable VT demonstrates the highest importance, with values of 0.28 in the TFCN model and 0.27 in the TFLSTM model. This prominence of VT surpasses that of other dynamic variables in contributing to the final predictions. Such findings indicate a robust relationship between VT and the estimation of $\dot{V}O_2$, thereby underscoring its critical significance in the realm of exercise physiology.

Other variables such as BF and WorkLoad maintain moderate importance across both models (ranging from 0.15 to 0.24). This indicates that while they may not be the strongest predictors, they still contribute significantly to the models' predictive capabilities.

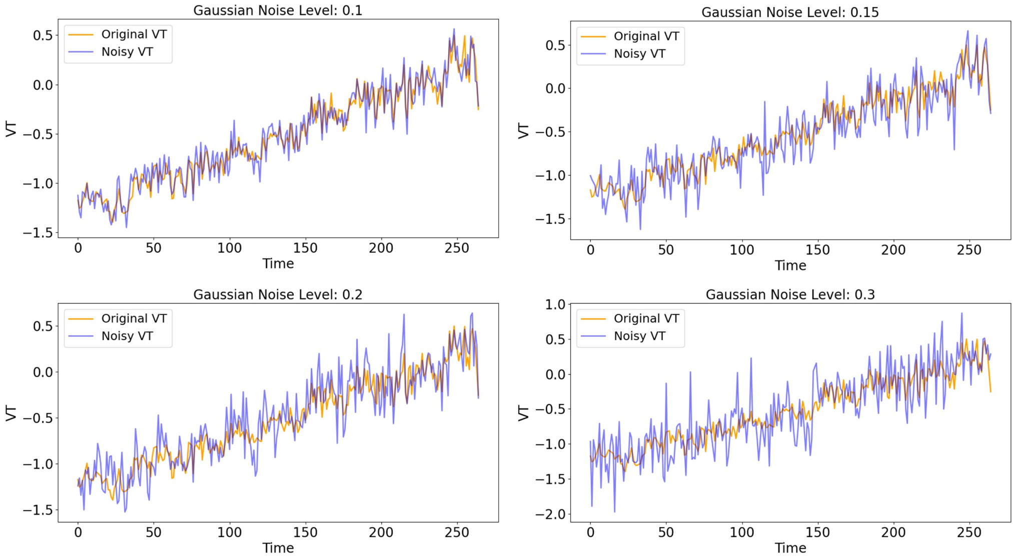

Fig. 6. The figure depicts the effects of four different types of noise on continuous data over time. The orange line represents the original continuous variable, while the line illustrates the variable after noise has been added.

In summary, the observations from analyzing the weights and importance values assigned to each variable in the model demonstrate its ability to provide quantifiable explanations for the model's performance. This level of interpretability is crucial, as it allows researchers to verify that the model is operating as expected based on domain knowledge.

4.4 Ablation Study

In this section, we examine the impact of anthropometric variables on the performance of our models, specifically the TFCN and TFLSTM. As illustrated in Table 3, the inclusion of anthropometric variables results in performance enhancements for both models.

The integration of these variables provides prior knowledge to the TFN models, leading to a reduction in RMSE of 0.04 L/min for the TFCN model and 0.03 L/min for the TFLSTM model. Additionally, the MAE is reduced by 0.05 L/min for both the TFCN model and 0.06 L/min for TFLSTM models. Correspondingly, the R^2 for both of the modes improves by 0.06. To substantiate these improvements, we conducted Wilcoxon signed-rank test on the R^2, RMSE, and MAE values of the models with and without anthropometric variables. Since we performed multiple tests for three different evaluation metrics, we applied Bonferroni correction to adjust the significance threshold, setting the adjusted threshold to $\alpha = \frac{0.05}{3} \approx 0.0167$. The null hypothesis was rejected only when the p-value was less than this adjusted threshold, leading to the updated reported p-values. The resulting p-values for the TFCN model were $(5.62e-04, 1.87e-05, 7.18e-05)$, and for the TFLSTM model, the p-values were $(2.41e-04, 3.85e-05, 1.03e-04)$. These results indicate that the observed changes in performance metrics are statistically significant, clearly demonstrating substantial performance improvements attributable to the inclusion of anthropometric variables. Furthermore, detailed graphical Fig. 7 reveal that the models incorporating anthropometric variables demonstrate performance metrics that are closer to the actual values, underscoring the relevance of these variables in enhancing model accuracy.

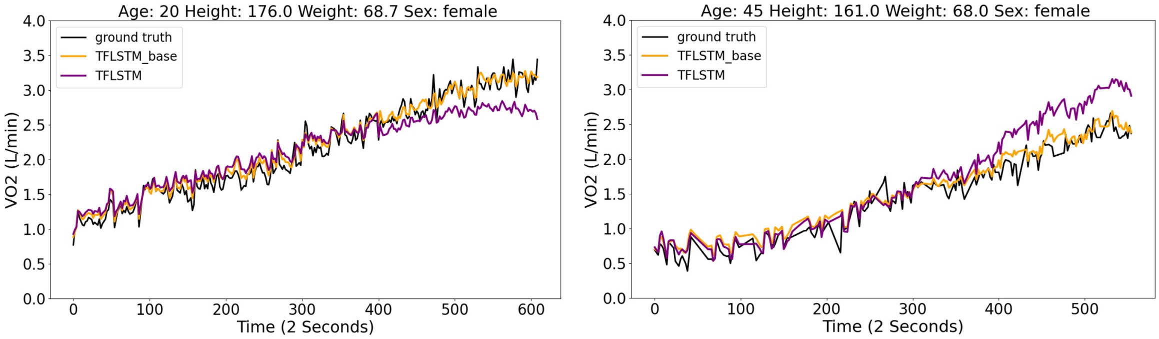

Fig. 7. This figure compares the performance of two TFN models (TFCN; TFLSTM), both with and without the inclusion of anthropometric variables. The black line depicts the actual $\dot{V}O_2$ values. Upon inspection, it is evident that the incorporation of these variables facilitates a more accurate learning of the temporal behavior of $\dot{V}O_2$.

Currently, the data utilized in our study is demonstrably of higher quality compared to data obtained from wearable devices. Consequently, the effect of this disparity on actual performance remains uncertain. Wearable sensors often produce data that is inherently noisy, and their accuracy can be compromised due to factors related to both equipment limitations and environmental conditions. To address this issue and assess the robustness of our model, we introduced noise to the dataset to simulate real-life scenarios.

Therefore, we introduced four distinct levels of Gaussian noise: 0.1, 0.15, 0.2, 0.3. The noise was generated from a normal distribution with a mean of zero, which ensured that the introduction of noise did not systematically bias the data. The results of this analysis are presented in Table 5, which outlines the impact of the added noise on the model's performance. Generally, when the noise level is less than or equal to 0.2, the performance of our model remains robust. Specifically, the R^2 values are consistently maintained within the range of 0.80 to 0.84 across different noise conditions in this interval. Additionally, both the RMSE and MAE metrics exhibit acceptable levels of prediction error, with RMSE values ranging from 0.11 to 0.12 and MAE values from 0.23 to 0.26. This consistency indicates that our model can effectively accommodate moderate noise levels while still delivering reliable predictions, which is particularly important for applications where data quality may fluctuate. However, as noise levels exceed 0.2, a noticeable decline in performance is observed, highlighting the need for careful consideration of the impact of noise during model evaluation and deployment.

Table 5. Model Performance Under Various Noise Conditions

	TFCN	+Gau(0.1)	+Gau(0.15)	+Gau(0.2)	+Gau(0.3)
R^2	0.92	0.82	0.82	0.80	0.76
RMSE	0.03	0.11	0.12	0.12	0.13
MAE	0.13	0.25	0.25	0.26	0.28

	TFCN	+Gau(0.1)	+Gau(0.15)	+Gau(0.2)	+Gau(0.3)
R^2	0.94	0.84	0.83	0.82	0.79
RMSE	0.02	0.11	0.11	0.11	0.13
MAE	0.11	0.23	0.24	0.25	0.27

5 DISCUSSION AND CONCLUSION

Our proposed model demonstrates the ability to take in CPET data of varying time lengths as input and produce accurate $\dot{V}O_2$ values along the time. Using a limited set of easy-to-measure features including VT, BF, HR, HRR, and basic anthropometrics, the model achieves state-of-the-art performance. A key finding of this work is the importance of incorporating anthropometric variables for precise $\dot{V}O_2$ estimation, highlighting the need to consider both physiological responses and individual characteristics for the future work. Meanwhile, the parsimony of inputs required by our model (many of which can be collected via portable devices) suggests promising applications for expanding $\dot{V}O_2$ monitoring beyond laboratory settings.

For example, integration with a wearable spirometry like MiniSpir to gather VT and BF alongside smartwatch collection of HR and HRR parameters could enable minimally-burdensome and affordable testing in field contexts. This has implications for increasing access to $\dot{V}O_2$ profiling, and enabling investigation of metabolic responses under real-world conditions rather than confined laboratory protocols.

While prior work has predominantly focused on enhancing the model performance, many of these existing methods can be characterized as "black-box" models that provide little insight into the relationships learned. In contrast, our work applied a variable selection mechanism allowing the model to explicitly determine the relative importance of different features. Such interpretability is valuable, as it provides useful insights for researchers in the domain.

Overall, our results demonstrate the potential for interpretable AI to leverage wearable-accessible indicators as a pathway to advancing non-exercise $\dot{V}O_2$ assessment. With further validation and interface with adjunct technologies, forecasting cardiorespiratory fitness from sparsely sampled signals collected during activities of daily life may become feasible. We believe the current work presents an exciting step toward more transparent, collaboratively optimized methods for $\dot{V}O_2$ analyses in medical and health domains.

6 DATA AND CODE AVAILABILITY

The data utilized in this study is owned by Manchester Metropolitan University (MMU) and is not publicly available. However, there is a formal agreement between KAUST and MMU allowing the use of this data for the purposes of this publication. Additionally, the code employed for the analysis is accessible on GitHub at the repository TFN-VO2. We encourage researchers to leverage these resources for further exploration and validation of our findings.

ACKNOWLEDGMENTS

Our Study has been approved by Institutional Biosafety and Bioethics Committee (IBEC) at King Abdullah University of Science and Technology (KAUST) under the IBEC registration number 23IBEC087.

REFERENCES

[1] David R Bassett and Edward T Howley. Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Medicine and science in sports and exercise*, 32(1):70–84, 2000.

[2] Matteo Bonato, Susanna Rampichini, Marco Ferrara, Stefano Benedini, Paola Sbriccoli, Giampiero Merati, Emerson Franchini, Antonio La Torre, et al. Aerobic training program for the enhancements of hr and vo2 off-kinetics in elite judo athletes. *J Sports Med Phys Fitness*, 55(11):1277–1284, 2015.

[3] Toshiyo Tamura. Wearable oxygen uptake and energy expenditure monitors. *Physiological Measurement*, 40(8):08TR01, 2019.

[4] Takeshi Ebihara, Kentaro Shimizu, Masahiro Ojima, Yohei Nakamura, Yumi Mitsuyama, Mitsuo Ohnishi, Hiroshi Ogura, and Takeshi Shimazu. Energy expenditure and oxygen uptake kinetics in critically ill elderly patients. *Journal of Parenteral and Enteral Nutrition*, 46(1):75–82, 2022.

[5] Rajeev Malhotra, Kristian Bakken, Emilia D'Elia, and Gregory D Lewis. Cardiopulmonary exercise testing in heart failure. *JACC: Heart Failure*, 4(8):607–616, 2016.

[6] Sergio Caravita, Ilaria Tanini, Lia Crotti, Claudia Baratto, Gianfranco Parati, Francesco Fattirolli, Iacopo Olivotto, and Franco Cecchi. Impaired cardiopulmonary test performance as a marker of early functional impairment in patients with anderson-fabry disease. *Journal of Cardiovascular Medicine and Cardiology*, pages 069–071, 11 2021.

[7] David C Nieman, Melanie D Austin, Dustin Dew, and Alan C Utter. Validity of cosmed's quark cpet mixing chamber system in evaluating energy metabolism during aerobic exercise in healthy male adults. *Research in Sports Medicine*, 21(2):136–145, 2013.

[8] COSMED. COSMED-Quark-CPET. (2024).

[9] Vacumed. CYCLUS2. (2024).

[10] Scott E Crouter, Samuel R LaMunion, Paul R Hibbing, Andrew S Kaplan, and David R Bassett Jr. Accuracy of the cosmed k5 portable calorimeter. *PLoS One*, 14(12):e0226290, 2019.

[11] Monika Petelczyc, Michał Kotlewski, Sven Bruhn, and Matthias Weippert. Maximal oxygen uptake prediction from submaximal bicycle ergometry using a differential model. *Scientific Reports*, 13(1):11289, 2023.

[12] Krzysztof Przednowek, Zbigniew Barabasz, Maria Zadarko-Domaradzka, Karolina H Przednowek, Edyta Nizioł-Babiarz, Maciej Huzarski, Klaudia Sibiga, Bartosz Dziadek, and Emilian Zadarko. Predictive modeling of vo2max based on 20 m shuttle run test for young healthy people. *Applied Sciences*, 8(11):2213, 2018.

[13] Fatih Abut, Mehmet Fatih Akay, and James George. A robust ensemble feature selector based on rank aggregation for developing new vo\textsubscript {2} max prediction models using support vector machines. *Turkish Journal of Electrical Engineering and Computer Sciences*, 27(5):3648–3664, 2019.

[14] Fatih Abut, Mehmet Fatih Akay, and James George. Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in biology and medicine*, 79:182–192, 2016.

[15] Bryson Carrier, Andrew Creer, Lauren R Williams, Timothy M Holmes, Brayden D Jolley, Siri Dahl, Elizabeth Weber, and Tyler Standifird. Validation of garmin fenix 3 hr fitness tracker biomechanics and metabolics (vo2max). *Journal for the Measurement of Physical Behaviour*, 3(4):331–337, 2020.

[16] Luyao Yang, Osama Amin, and Basem Shihada. Intelligent wearable systems: Opportunities and challenges in health and sports. *ACM Computing Surveys*, 56(7):1–42, 2024.

[17] Carmen Spaccarotella, Alberto Polimeni, Cinzia Mancuso, Girolamo Pelaia, Giovanni Esposito, and Ciro Indolfi. Assessment of non-invasive measurements of oxygen saturation and heart rate with an apple smartwatch: comparison with a standard pulse oximeter. *Journal of clinical medicine*, 11(6):1467, 2022.

[18] Zachi I Attia, David M Harmon, Jennifer Dugan, Lukas Manka, Francisco Lopez-Jimenez, Amir Lerman, Konstantinos C Siontis, Peter A Noseworthy, Xiaoxi Yao, Eric W Klavetter, et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nature medicine*, 28(12):2497–2503, 2022.

[19] Jeffrey Montes, John C Young, Richard Tandy, and James W Navalta. Reliability and validation of the hexoskin wearable bio-collection device during walking conditions. *International journal of exercise science*, 11(7):806, 2018.

[20] Andrea Zignoli, Alessandro Fornasiero, Matteo Ragni, Barbara Pellegrini, Federico Schena, Francesco Biral, and Paul B Laursen. Estimating an individual's oxygen uptake during cycling exercise with a recurrent neural network trained from easy-to-obtain inputs: A pilot study. *PLoS One*, 15(3):e0229466, 2020.

[21] Robert Amelard, Eric T Hedge, and Richard L Hughson. Temporal convolutional networks predict dynamic oxygen uptake response from wearable sensors across exercise intensities. *NPJ digital medicine*, 4(1):156, 2021.

[22] Marco Altini, Julien Penders, and Oliver Amft. Estimating oxygen uptake during nonsteady-state activities and transitions using wearable sensors. *IEEE journal of biomedical and health informatics*, 20(2):469–475, 2015.

[23] Md Mobashir Hasan Shandhi, William H Bartlett, James Alex Heller, Mozziyar Etemadi, Aaron Young, Thomas Plötz, and Omer T Inan. Estimation of instantaneous oxygen uptake during exercise and daily activities using a wearable cardio-electromechanical and environmental sensor. *IEEE journal of biomedical and health informatics*, 25(3):634–646, 2020.

[24] Atiqa Ashfaq, Neil Cronin, and Philipp Müller. Recent advances in machine learning for maximal oxygen uptake (vo2 max) prediction: A review. *Informatics in Medicine Unlocked*, 28:100863, 2022.

[25] Maria Cecília Moraes Frade, Thomas Beltrame, Mariana de Oliveira Gois, Allan Pinto, Silvia Cristina Garcia de Moura Tonello, Ricardo da Silva Torres, and Aparecida Maria Catai. Toward characterizing cardiovascular fitness using machine learning based on unobtrusive data. *Plos one*, 18(3):e0282398, 2023.

[26] Pinaki Chatterjee, Alok K Banerjee, Paulomi Das, and Parimal Debnath. A regression equation for the estimation of maximum oxygen uptake in nepalese adult females. *Asian journal of sports medicine*, 1(1):41, 2010.

[27] Amit Bandyopadhyay. Validity of 20 meter multi-stage shuttle run test for estimation of maximum oxygen uptake in male university students. *Indian J Physiol Pharmacol*, 55(3):221–226, 2011.

[28] Gustavo Silva, Nórton Luis Oliveira, Luísa Aires, Jorge Mota, José Oliveira, and José Carlos Ribeiro. Calculation and validation of models for estimating vo2max from the 20-m shuttle run test in children and adolescents. *Arch Exerc Health Dis*, 3(1-2):145–152, 2012.

[29] Steven W Su, Lu Wang, Branko G Celler, and Andrey V Savkin. Oxygen uptake estimation in humans during exercise using a hammerstein model. *Annals of biomedical engineering*, 35:1898–1906, 2007.

[30] Andrew J Cook, Ben Ng, Gaetano D Gargiulo, Diane Hindmarsh, Mark Pitney, Torsten Lehmann, and Tara Julia Hamilton. Instantaneous vo2 from a wearable device. *Medical Engineering & Physics*, 52:41–48, 2018.

[31] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

[32] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

[33] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32, 2019.

[34] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

[35] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[36] Gary J Balady, Ross Arena, Kathy Sietsema, Jonathan Myers, Lola Coke, Gerald F Fletcher, Daniel Forman, Barry Franklin, Marco Guazzi, Martha Gulati, et al. Clinician's guide to cardiopulmonary exercise testing in adults: a scientific statement from the american heart association. *Circulation*, 122(2):191–225, 2010.

[37] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

[38] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv*, abs/1803.01271, 2018.

[39] JEROME L Fleg and EDWARD G Lakatta. Role of muscle loss in the age-associated reduction in vo2 max. *Journal of applied physiology*, 65(3):1147–1151, 1988.

[40] Š Šprynarová, J Pařizková, and V Bunc. Relationships between body dimensions and resting and working oxygen consumption in boys aged 11 to 18 years. *European journal of applied physiology and occupational physiology*, 56(6):725–736, 1987.

[41] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21:137–146, 2011.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009