Redefining and Interpret Instantaneous Oxygen Uptake Estimation with Temporal Fusion Convolutional Networks

Luyao Yang, *Student Member, IEEE*, Osama Amin *Senior Member, IEEE*, Azmy Faisal, and Basem Shihada, *Senior Member, IEEE*

***Abstract*— Oxygen consumption (VO_2) is a well-established clinical and physiological marker of cardiorespiratory functionality and exercise capacity. Despite this, VO_2 monitoring is predominantly restricted to expensive equipment and specialized laboratory environments, thus limiting its broader application. To address this limitation, our study introduces an end-to-end Temporal Fusion Convolutional Network (TFCN). This model leverages easily accessible physiological parameters, such as heart rate (HR), heart rate reserve (HRR), minute ventilation (VE), tidal volume (VT), and breathing frequency (BF), to accurately predict VO_2 dynamics. We derive these variables from cardiopulmonary exercise testing (CPET) data, which was collected from a diverse group of 58 adults using a COSMED system (41 males, 17 females; 38 healthy, 20 smokers; average age: 30±15 years; height: 1.75±0.2m; weight: 90±30kg; VO_2 max: 42±6 L/min). Each participant was subjected to an incremental exercise protocol, facilitating a comprehensive exploration of VO_2 dynamics. Ultimately, our base TFCN model demonstrated a robust performance with an RMSE of 0.30 L/min and an R_2 of 0.84. Notably, the refined TFCN model further enhanced these results, achieving an improved RMSE of 0.23 L/min and an R_2 of 0.92. Our study establishes the feasibility of predicting VO_2 dynamics using low-cost, readily available variables outside of a laboratory setting. Additionally, we examined the weight of each input variable for a comprehensive interpretation of the final VO_2 predictions. Our study illustrates the potential of our model in delivering highly accurate VO_2 predictions in non-laboratory settings, enhancing its reliability and interpretability in potential applications.**

***Index Terms*— Oxygen Uptake, Deep learning, VO_2 estimation**

I. Introduction

OXYGEN uptake (VO_2) is the standard metric used to gauge the shifts in an individual's absorption, transportation, and utilization of oxygen as a response to variations in exercise intensity. This indicator provides a quantifiable measure of how efficiently the body adapts to the oxygen demands imposed by physical activity. VO_2 measurement is important due to numerous reasons. Studies measure the instaneous VO_2 to estimate energy expenditure (EE) that improves the nutrition management of elderly patients who need to enter intensive care units frequently [1], [2]. It also serves as a biomarker in the medical research field, employed by scientists to quantify the progression and current status of a patient's illness such as heart failure and even rare diseases [3], [4].

VO_2 max, obtained from instaneous VO_2, representing the peak VO_2 an individual can absorb during incremental exercise. VO_2 max gives a benchmark of the athlete's current level of aerobic fitness, providing a quantitive indicator of athlete's aerobic capacity. Coaches could utilize it to design tailored training plan that cater to each athlete [5].

The most commonly used and accurate way to measure the VO_2 is cardiopulmonary exercise testing (CPET), which requires the people wear a mask or mouthpiece that connected to a metabolic cart. Then the machine captures and measures the volume and gas concentraions of inhaled and exhaled air. However, the machine is huge that must be tested by professional experimenters to operate in lab environment. Therefore, some commercial products have emerged to propose more convenient alternative detection methods. Enterprises such as COSMED and VacuMed have developed portable devices that enable outdoor testing capabilities. However, the expensive price and the requisite use of a facial mask that cling to the head introduces limitation and unconvenience. Hence, its application for extended durations during routine daily activities remains constrained [6].

Based on these limitation, some researches try to measure the VO_2 max by establishing a multivariate equation based on the input such as age, sex, weight and height of each individual [7]–[10]. In addition, commercial products such as Garmin and Samsung smartwatches also measure the VO2 max [11]. Nonetheless, these methodologies are limited to the computation of singular VO_2 values and their models lack generalization to diverse, novel populations. Furthermore, the high degree of error associated with these methods compromises their applicability in realistic, everyday scenarios [8], [11].

Recent advance in wearable devices and artificial intelli-

Manuscript received Sep 15, 2023; This work was supported in part by ....

Luyao Yang, Osama Amin, and Basem Shihada are with the King Abdullah University of Science and Technology, Thuwal, KSA (e-mail: luyao.yang@kaust.edu.sa; osama.admin@kaust.edu.sa; basem.shihada@kaust.edu.sa).

Azmy Faisal is with the Manchester Metropoliton University, Manchester, UK (e-mail: azmy.faisal@mmu.ac.uk)

gence forged a new way for researchers. Studies have shown the smartwatches could provide high-accuracy indicators such as heart rate (HR), Electrocardiogram (ECG), oxygen saturation (SpO_2) and so on [12], [13]. Pertaining to respiratory parameters, smart garments such as HEXOSKIN offer real-time measurements of indices like Minute Ventilation (VE) and Breathing Frequency (BF). Accordingly, the accuracy and validity of these parameter measurements have been substantiated by research [14]. Thus, some methods utilize the AI methods to predict the VO_2 with inputs from these low-cost and easy-to-obtain variables such as HR, VE, BF, etc [15]–[18]. However, these methods only focus on the continuous variables derived from wearables, representing data that alters over the time. While these anthropometric variables which are not used by these methods, which could potentially harbor a wealth of information, such as men typically exhibiting higher VO_2 values than women. Besides, they cannot provide explainable insights to explain which input variable is most important.

Thus, we propose a novel deep learning model, which combines both variable variables and anthropometric variables to predict the instaneous VO_2. Our paper has the following contributions:

1) We design a temporal fusion convolutional network (TFCN) to provide the insights that we could use low-cost parameters to estimate the instaneous VO_2 in a non-invaisve and long-time way.
2) We use more data(58), and different time span for each sample to achieve the highest accuracy among the existing methods.
3) Compared with other methods, we incoporate the anthropometric variables such as gender, age, weight, height and BMI as the enhancer of our model.
4) We are the first model that consider temporal behavior including the time and workload.
5) We provide the explainable weights of each inputs variables, which provides insights for the future research.

II. Related Works

Within the corpus of related literature, numerous studies have focused on predicting singular VO_2 max values, employing a range of statistical techniques and machine learning methodologies [19]. However, only a handful of these studies have ventured into the realm of instantaneous VO_2 prediction [16]–[18]. Now our work intends to fill this gap by concentrating on the prediction of instantaneous VO_2.

A. VO_2 max prediction

Due to the importance of VO_2 max, the early work adopt the questionnaires to collect the basic information of each individual. And then use mathematical modeling or machine learning methods to build a predictive model. They utilize the anthropometric variables such as age, height, BMI, weight, gender, HR max to calculate the VO_2 max [10]. Table I shows the works and their related methods and input variables.

TABLE I
SUMMARY OF VO_2 MAX PREDICTION METHODS

Study	Methods	Input Variables
Petelczyc et al. 2023 [7]	Differential model	Gender, age, HR, HRmax, Workload
Abut et al. 2019 [9]	SVM, GRNN, SDT	Gender, age, height, weight, HRmax, time, HR
Przednowek et al. 2018 [8]	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 [10]	SVM, MLP	Gender, age, MX-HR, SM-ES, Q-PFA

B. Instantaneous VO_2 prediction

Predicting instantaneous VO_2 presents more substantial challenges compared to the estimation of a singular maximal VO_2 value. This complexity arises from the necessity to extract and learn a greater volume of temporal features for instantaneous VO_2 prediction. Furthermore, the instantaneous VO_2 alterations across different exercises and diverse individuals further complicates the prediction process. While VO_2 max is determined under peak exertion within a controlled test, the measurement of instantaneous VO_2 mandates continuous monitoring during the course of exercise, posing significant technical challenges. In this section, we concentrate on reviewing recent methodologies aimed at predicting instantaneous VO_2. A summary of these methods is provided in Table II. Finally, in response to the identified limitations of these existing approaches, we propose our own study.

The initial attempts to quantify dynamic VO_2 can be traced back to the Su et al's work published in the year 2007. They used the Support Vector Regression (SVR) methods to establish a VO_2 prediction model learned from the Pseudo-Random Binary Sequence (PRBS) signal in the running activity [20]. However, the limitation lies in its sole reliance on treadmill speed for predicting VO_2, without considering inter-individual differences. In other words, the same treadmill speed could yield different VO_2 values across individuals, reflecting the unique physiological responses of each person.

With the ongoing advancements in wearable technology, which can capture intricate physical parameters from the human body, there has been a surge in studies that seek to apply machine learning and statistical methods to data gathered from these wearable devices [15]–[18], [21].

Altini et al. were pioneers in using accelerometer (ACC) and HR sensor data to estimate nonsteady-state VO_2 during transitions between daily activities, including lying, sitting, walking, biking, and running. By leveraging Support Vector Machine (SVM) techniques, they developed a range of models specifically tailored to predict VO_2 for different activities and their transitions [17]. Nonethelessness, the necessity to create a new model for each state is a rather cumbersome process and lacks universality. Then Cook et al. designed a algorithm

TABLE II
SUMMARY OF METHODS FOR PREDICTING INSTANTANEOUS VO_2

Study	Data size	Training data	Testing data	Methods	Model inputs	Device	Protocols or exercise
Su et al. 2007 [20]	6	6	6	SVR	PRBS and speed	Treadmill	Use PRBS to control the treadmill protocol
Altini et al. 2015 [17]	22	21	1(LOOCV)	SVM	ACC, HR, anthropometric features	ECG necklace, ACC	Lying, sedentary, dynamic, walking, biking, running
Cook et al. 2018 [21]	42	28	14	IAA	ECG, ACC, HR	DREEM, COSMED K4b^2	Bruce protocol
Zignoli et al. 2020 [15]	7	7	7	LSTM	HR, RF, P, ω	power meter, COSMED	Arbitrary protocols, Wingate test
Shandhi et al. 2020 [18]	17	16	1(LOOCV)	XGBoost	SCG, ECG, AP	Custom-built wearable patch	Treadmill protocol, outside protocol
Amelard et al. 2021 [16]	22	17	5	TCN	HR, HRR, RF, VE	Smart shirt	One ramp-incremental, PRBS protocol
Our study	58	40	18	TFT	Gender, age, height, weight, BMI, HR, HRR, VE, VT	COSMED	Incremental exercise

Altini et al. 2015 [17] and Shandhi et al. 2020 [18], they use the Leave-One-Out Validation (LOOV) for testing; Zignoli et al. 2020 [15] trained their model on two protocols per person and tested it on a different protocol for each.

combining HR and the integral of absolute acceleration (IAA) to detect instantaneous VO_2 [21].

To make the model more applicable in different exercises, Shandhi et al. used a built wearable patch placed on the mid-sternum to collect seismocardiography (SCG), electrocardiogram (ECG), and atmospheric pressure (AP) signals from 17 adults using both inside treadmill protocols and outside protocols [18]. 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.

In addition, neural networks, especially deep learning models, 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 predict VO_2. [15]. 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, they used a power meter mounted on the bicycle instead of a wearable sensor as the input, and thus can only be applicable for 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. is the first work using deep learning mode temporal convolutional network (TCN) to predict VO_2 based on the smartshirt inputs. 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 temporal convolutional network (TCN) on 17 adults and test the model on the rest 5 adults [16]. This work shows the great power of deep learning in prediction the instantaneous VO_2. However, the author ignored the physical and anthropometric features of each individuals, which are very important in deciding the VO_2. Simultaneously, the model falls short in providing explanations for its input factors and lacks extensive validation to ensure its generalization capabilities.Furthermore, the age variance among most participants in the study is less than 10 years, weight variation is within 20kg, and height disparity is less than 10cm. This suggests that the participant group is homogeneous.

Therefore, based on these limitation, as shown in Table II, we collect a larger dataset with 58 participants (41 males, 17 females; 38 healthy, 20 smokers; average age: 30±15 years; height: 1.75±0.2m; weight: 90±30kg; VO_2 max: 42±6 L/min). We aim to combine both anthropometric features (such as gender, age, height, weight, Body Mass Index (BMI)) and the temporal features that can be obtained from wearables (such as WorkLoad, HR, HRR, BF, VE, VT) into a deep learning model to accurately predict VO_2. We validate that our model can perform well and generalized on different separate groups including the young and old, smoker and healthy groups.

III. METHODS

VO_2 kinetics demonstrate an organized but intricate temporal pattern in response to exercise. Thus extracting temporal information poses a significant challenge. Among prior studies in Section 2, only Amelard et al. attempted to harness a TCN to capture temporal dynamics [16]. Although TCN obtains a good performance, it still can not give us som useful information that which variables matters a lot to our predictions. To enhance both the interpretability and performance, we have developed a novel model called Temporal Fusion Convolutional Network (TFCN), as illustrated in Fig. 1. This model is built upon two sections including the feature embedding and temporal extraction part. Inspired by [22], we built a feature embedding module that converts the anthropometric and dynamic inputs into the input embeddings, and also includes the importance of each variable. And then we established a temporal extraction module that learn the temporal features of the inputs sequences. In the next sections, we will have a clearer understanding about the method implementation and model structure.

A. Overview of the Method

Given a time-series input sequence $x = x_1, x_2, ..., x_t, ..., x_T \in \mathbf{R^n}$. For each input sequence, it has T timesteps. At the current time t, x_t represents the input variables values at time t, which is a n-dimension vector. n equals to the number of variables we used including the dynamic and anthropometric variables (in this case, the $n = 11$). The aim of the model is to predict the $VO_{2(t)}$ value given the historical time-series $x_1, x_2, ..., x_t$. After concurrently processing all input sequence with the T timesteps, the model would predict the according temporal sequence $\hat{VO}_2 : y = y_1, y_2, y_3, ..., y_t, ..., y_T \in \mathbf{R}$, and y_t represents the predicted $\hat{VO}_{2(t)}$ value at current t timestep.

$$\hat{VO}_{2(t)} = y_t = Model(x_1, x_2, x_3, ..., x_t) \tag{1}$$

Overall, as the equation (1) shows, the model is trained to minimize the difference between its predicted $\hat{VO}_2$ values and the true VO_2 values.

B. Model Structure

Our Temporal Fusion Convolutional Network (TFCN) model, depicted in Fig. 1, is primarily composed of three main components: the feature selection part, the temporal extraction part and the output generation part.

1) Feature Selection: Drawing inspiration from prior work in interpretable time-series modeling [22], this component comprises two independently parameterized variable selection networks based on Gated Residual Network (GRN) - one dedicated to static anthropometric features, the other for dynamical signals. Each network embeds its respective input variables into lower-dimensional spaces, allowing the model to quantify the predictive importance of each original variable. Specifically, the anthropometric network projects attributes like age, sex and body dimensions, while the dynamic network encodes time-series measurements including VE, VT, BF, HR and HRR.

Crucially, the variable selection procedure assigns a numerical importance weight to each embedded variable representation. These weights indicate the relative contribution of the original input features in informing the overall model's predictions. For example, variables assigned higher weights can be interpreted as carrying more predictive information regarding future VO_2 levels. This module enhances model transparency while facilitating discovery of key determinants among the input variables.

2) Temporal Extraction: A crucial component of modeling temporal data is adequately capturing the temporal dependencies and dynamics within the inputs cariables. To address this, our model incorporates a specialized Temporal Extraction Part based on temporal convolutional networks (TCNs).

TCNs are particularly well-suited for extracting meaningful temporal representations from sequential data due to their dilated causal convolutions which allow the model to learn patterns over a wide effective historical context. In our implementation, the Temporal Extraction Part contains a 3-layer 1D TCN architecture. Each layer doubles the receptive field size through an exponentially increasing dilation factor. This recursive structure enables the network to identify patterns spanning both short-term fluctuations and long-range trends over the entire input sequence history.

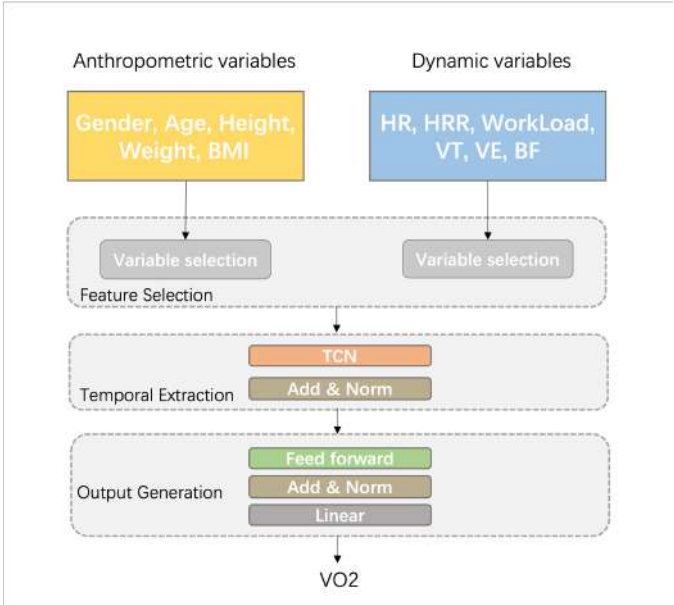

Fig. 1. Schematic representation of the Temporal Fusion Convolutional Network (TFCN) model. The model is divided into three main sections: Variable Selection uses two networks to process anthropometric and dynamic variables from the input data. Temporal Extraction uses a Temporal Convolutional Network (TCN) and a self-attention mechanism to understand patterns and importance over time. The final part, Output Generation, uses a feed-forward network to predict the $\hat{\boldsymbol{VO_2}}$ values.

By extracting multi-scale temporal features from the embedded input variables, the TCN aids downstream predictive and explanatory modeling. Its dilated architecture also helps address the variable-length nature of real-world physiological time-series.

3) Output Generation: After the Variable Selection and Temporal Extraction parts process and transform the input data, the extracted high-level feature representations are passed to the final Feed-Forward Network (FFN). The FFN contains multiple fully-connected layers that act as a regressor, projecting the encoded multi-dimensional feature space onto the target variable space of $\hat{VO}_2$ values. Specifically, the temporal patterns learned by the earlier components are mapped through the deep set of nonlinear activations within the FFN. By fusing all available input information, the FFN performs the classification task of predicting $\hat{VO}_2$ using the linear layer. This last step effectively ties together all the learnings and transformations from the previous stages to produce the output that we are interested in predicting.

Altogether, the inclusion of the FFN completes the end-to-end predictive modeling pipeline. It ties together all the transformations performed by the preceding interpretable variable selection and representation learning stages to distill them into informative VO_2 forecasts. The full TFCN architecture is thus highly effective for both prediction and interpretability.

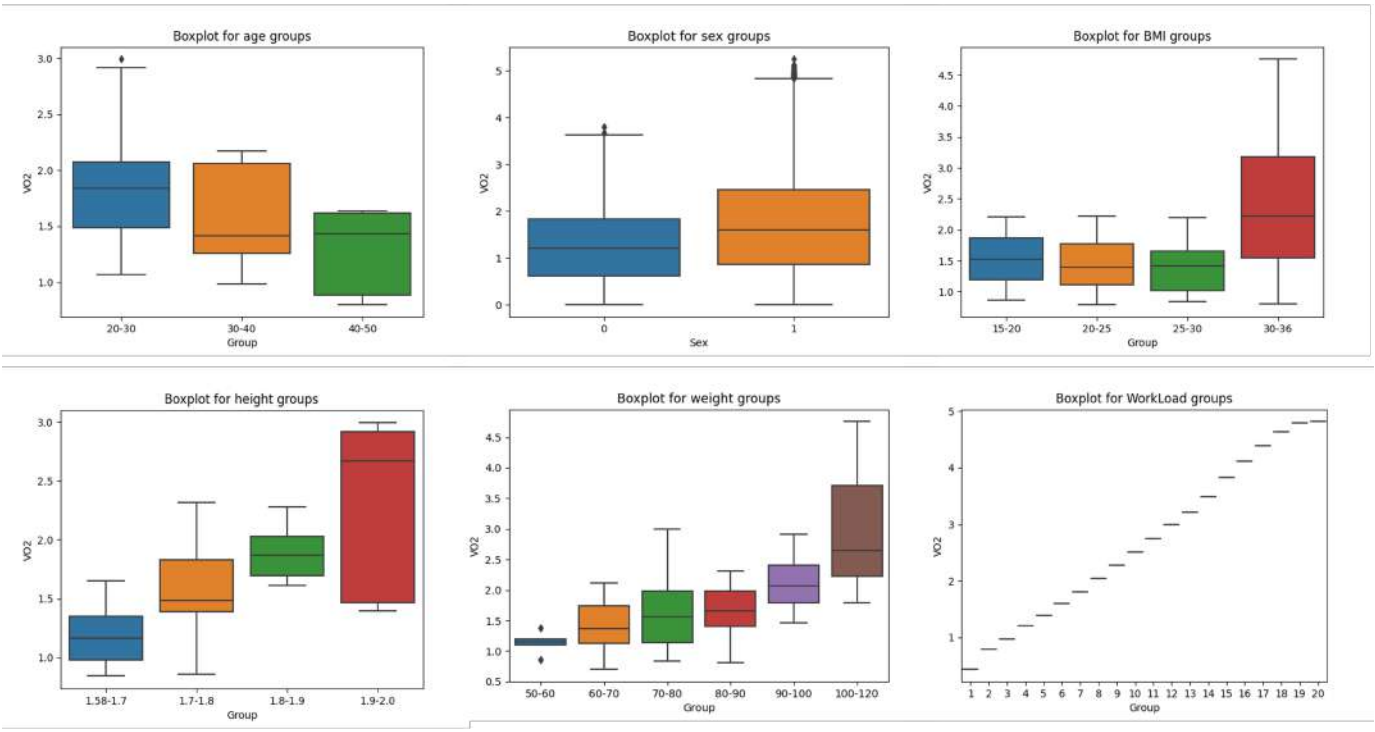

Fig. 2. For each specific anthropometric variable, the mean VO_2 value was calculated within each group. These calculations revealed noticeable disparities between the groups, indicating the distinct influence of each variable on VO_2 values.

C. Loss function

Quantile loss function, is often used in time series prediction when we are interested in predicting an interval estimate (quantile) instead of a point estimate. Specifically, the quantile loss function is designed to estimate the τ-th quantile of the conditional distribution of the response variable. By varying τ, the model can provide a different level of confidence in its estimates, which can be very useful when dealing with uncertainty and variability in time series data. The quantile loss L i-th timestep can be represented as the equation (2). For the total loss, the loss can be calculated by summing all quantile loss, N is the total number of timesteps, $\tau_{\max}$ is the total number of quantiles, and the double sum is taken over all timesteps i and all quantiles τ as equation (3) shows.

$$L_\tau(VO_{2(i)}, \hat{VO_{2(i)}}) = \tau * max(VO_{2(i)} - \hat{VO_{2(i)}}, 0) + (1-\tau) * max(\hat{VO_{2(i)}} - VO_{2(i)}, 0) \quad (2)$$

$$L_\tau(VO_2, \hat{VO_2}) = \sum_{\tau=1}^{\tau_{max}} \sum_{i=1}^{N} \frac{L_\tau(VO_2(i), \hat{VO_2}(i))}{N\tau_{max}} \quad (3)$$

Quantile loss allows the model to predict a range of possible outcomes, which can provide more information than a simple point prediction. We use three various percentiles (e.g. 10th, 50th and 90th) at each timestep, which means the $\tau = 0.1, 0.5, 0.9$, and finally we use the percentile the 50th as our predictions.

D. Data Availability

1) Data Collection: Fifty-eight adults (41 males, 17 females; 38 healthy, 20 smokers; average age: 30±15 years; height: 1.75±0.2m; weight: 90±30kg; VO_2 max: 42±6 L/min) volunteered to be measured by COSMED in this study. The experiment was under ...

2) Data Preprocessing: In the initial stages of our analysis, we began by procuring the raw data, which was subsequently subjected to a visual examination to identify potential outliers. Upon identification, these outliers were manually 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:

$$y_i = \frac{1}{k} \sum_{j=0}^{k-1} x_{i-j} \quad (4)$$

Unlike previous methodologies that collect data within a

fixed temporal span [16], [18], 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 VO_{2max}.

Our analytical framework leverages the sliding window to accommodate these differing time lengths. Let us denote the length of our sliding window as e, and the sequence length of a specific individual as S_i, with i serving as the unique identifier for the individual.

The window initiates at the sequence's commencement, encapsulating the initial e steps. Subsequently, the window shifts one step to the right, a process which continues iteratively. This procedure yields $S_i - e + 1$ samples for an individual i. In this way, not only augments our dataset but also effectively manages the inherent variability in sequence lengths across our dataset.

Finally, we employed a stratified sampling approach for our data split. In this method, 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.

3) Data Distribution: The distribution of the entire dataset is presented in Fig. 2, which illustrates the mean distribution of the variable VO_2 with respect to various parameters including gender, age, height, weight, BMI. The figure provides a visual representation of these anthropometric data characteristics, facilitating a more comprehensive understanding of the dataset.

The proposition that VO_2 is influenced by age, attributed to the decrease in metabolically active tissue associated with aging, has been in consideration since 1988 [23]. Concurrently, sports scientists began to recognize the effect of factors such as gender, weight, and BMI on VO_2 [24]. This is intuitively comprehensible, as a larger body necessitates a greater oxygen supply for its functioning. Furthermore, it was observed that men, on average, exhibit higher VO2 max values than women. Consequently, these anthropometric parameters were utilized into early methodologies for constructing estimations of VO_2. Given the distinct disparities observed among different groups in Fig. 2, we decided to incorporate anthropometric variables as prior knowledge into our model. This integration is designed to improve the performance of the model by capturing the influence of these factors.

E. Evaluation Metrics

The evaluation metrics we use are Root Mean Square Error (RMSE) and the R-squared (R^2). Suppose we have the true value in the i-th timestep of $VO_{2(i)}$ and the predicted value $\hat{VO}_{2(i)}$, the Root Mean Square Error (RMSE) could be calculated by equation (5).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(VO_{2(i)} - \hat{VO}_{2(i)}} \tag{5}$$

R^2 provides a quantitative measure of how well the predicted $\hat{VO}_2$ values from a model align with the actual values VO_2. R^2 is defined as the proportion of the variance in the dependent variable (in this case, the true VO_2) that is predictable from the independent variable(s) (in this case, the predicted $\hat{VO}_2$). 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 alculate the mean of the true VO_2 values, denoted as $\bar{VO}_2$:

$$\bar{VO}_2 = \frac{1}{n}\sum_{i=1}^{n} VO_{2(i)} \tag{6}$$

Compute the total sum of squares (TSS), which is the sum of squares of the difference between the true VO_2 value and the mean of true VO_2 values. It quantifies the total variance in the data:

$$TSS = \sum_{i=1}^{n}(VO_{2(i)} - \bar{VO}_2)^2 \tag{7}$$

Then calculate the residual sum of squares (RSS), which is the sum of squares of the difference between the true VO_2 value and the predicted $\hat{VO}_2$ value. It quantifies the variance left unexplained by the model:

$$RSS = \sum_{i=1}^{n}(VO_{2(i)} - \hat{VO}_{2(i)})^2 \tag{8}$$

Finally, based on the (7) and (8), we could calculate R^2 using the formula:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{9}$$

IV. Results

A. Comparison with other methods

In prior research, various studies have explored the utilization of LSTM and TCN for the prediction of VO2 [15], [16]. However, a conspicuous gap in the literature is the lack of available code for these predictive models, thereby limiting comparative analysis of their performance. In this study, we implemented the LSTM and TCN for the prediction of instantaneous VO2 using two different inputs sets: dynamic-variables only, dynamic variables and anthropometric variables. For each model, 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.

The performances of these models were compared in terms of their RMSE and Coefficient of Determination R^2 values, as shown in Table III. Among the various methods tested, our TFCN model exhibited state-of-the-art performance, achieving

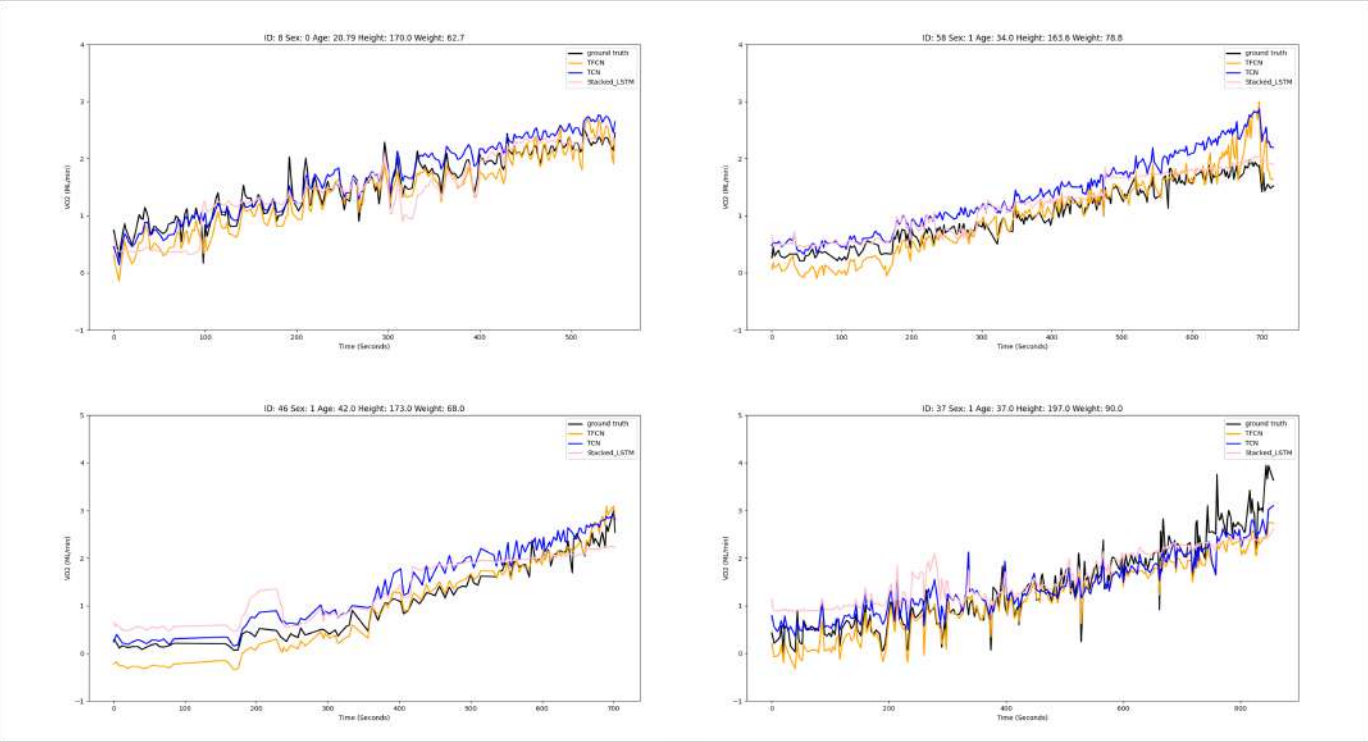

Fig. 3. We present the selected optimal scenarios across three age groups (20-30, 30-40, 40-50). The TFCN (yellow) demonstrates superior regression performance to the true VO_2 values (black) compared to the other two models (TCN: purple; LSTM: pink). It's noteworthy that all three models were supplied with identical inputs, encompassing both dynamic and anthropometric variables.

TABLE III
COMPARISON OF $RMSE_{(L/min)}$ AND R^2 FOR VO_2)

Methods	R^2	RMSE
Base Stacked LSTM	0.837	0.248
LSTM	0.690	0.300
Base TCN	0.626	0.402
TCN	0.802	0.273
Base TFCN	0.840	0.303
TFCN	**0.916**	**0.234**

TFCN model achieve the best performance than other methods, and the inputs including the anthropometric variables (gender, age, weight, height, BMI, workload) and the dynamic variables (HR, HRR, VE, VT, BF)

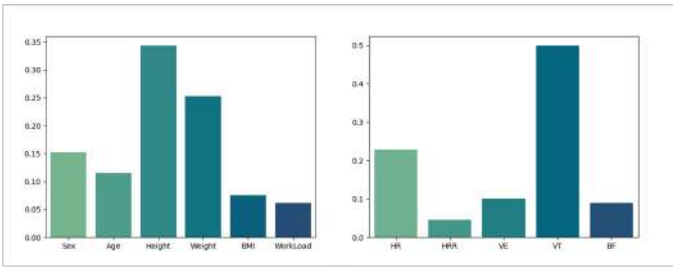

Fig. 4. This figure shows the importance of anthroupometric variables (left) and dynamic variables (right) separately.

the lowest RMSE and highest R^2 values. And we tested our all the models on the 18 testing files, Fig. 3 shows the food scenarios in three different age groups, TFCN shows the robustness compared with other modesl.

Apart from superior accuracy, our model offers explanatory insights into anthropometric and dynamic variables, the weights are obtained by averaging each sample importance as demonstrated in Table IV. The importance value assigned to each input variable serves as a measure of its significance to the final VO_2 prediction. As shown in Section methods, we quantify our variable importance by analyzing the variable selection weights which were calculated in the module feature selection.

An examination of the anthropometric variables reveals that height and weight each contribute nearly equal importance (approximately 0.5) to the accuracy of the predictions, as depicted in Figure 4. Among dynamic variables, VT holds the most importance, outweighing other variables in obtaining the final predictions. It is notable that the HRR variable was assigned the least importance compared to other variables in the time series prediction model. This finding can be reasonably explained by considering how the HRR metric is calculated. Specifically, the HRR is derived from three key heart rate measurements: $HR_{activity}$, the heart rate value recorded during each activity; HR_{rest}, the average heart rate during rest periods; and HR_{max}, the maximum heart rate observed. As the HRR is effectively an aggregated and normalized metric computed from these underlying HR variables, it contains less unique predictive power on its own compared to HR variable.

The observations from analyzing the weights and predictive importance values assigned to each variable in the model

demonstrate its ability to provide quantifiable explanations for predictions. This level of interpretability is crucial, as it allows researchers to verify that the model is operating as expected based on domain knowledge.

TABLE IV
IMPORTANCE

Anthropometric variable	Importance	Dynamic variable	Importance
Height	0.34	VT	0.5
Weight	0.25	HR	0.23
Sex	0.15	VE	0.10
Age	0.11	BF	0.09
BMI	0.07	HRR	0.04
WorkLoad	0.06	-	-

This table represents the importance of each variable to get the final VO_2 prediction

B. Ablation Study

As previously discussed and demonstrated in Table III and Table IV, the inclusion of anthropometric variables significantly improves the performance of our TCFN model and the TCN model. This enhancement is further visualized in Fig. 5, where each model's performance is compared both with and without the inclusion of anthropometric variables.

Incorporating these variables imparts prior knowledge to both the TCFN and the TCN models, leading to a reduction in RMSE by $0.07L/min$ and $0.13L/min$, respectively. This is a clear indication of the substantial performance improvement these variables bring to the TCN model.

Contrastingly, an increase in the number of variables results in a decrease in the performance of the LSTM model. This could potentially be due to the simpler structure of the LSTM model, which may inhibit its ability to effectively learn and process the information contained within the incorporated variables.

In sum, these findings substantiate, to a considerable degree, that both our model and the TCN model have successfully integrated and learned from the prior knowledge provided by the anthropometric variables.

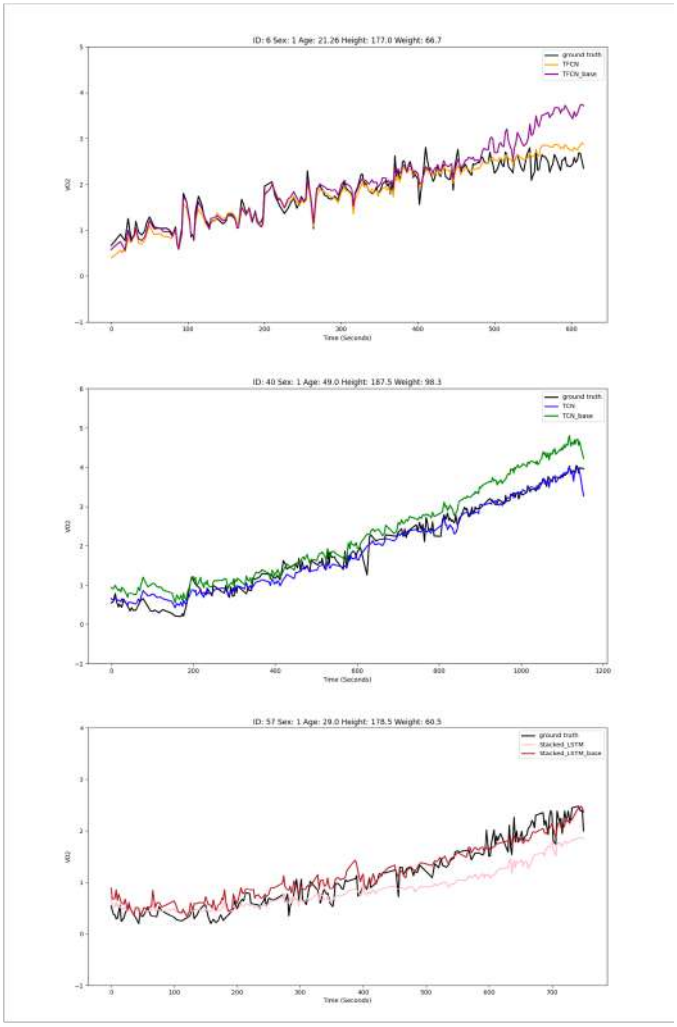

Fig. 5. This figure compares the performance of three models, both with and without the inclusion of anthropometric variables. The black line depicts the actual VO2 values. Upon inspection, it is evident that for both the TFCN and TCN, the incorporation of these variables facilitates a more accurate learning of the temporal behavior of VO_2. Conversely, for the LSTM model, the addition of these variables appears to have an adverse effect on its performance

V. Conclusion And Discussion

Our proposed model demonstrates the ability to take in CPET data of varying time lengths as input and produce accurate VO_2 values along the time. Using a limited set of easy-to-measure features including VE, VT, BF, HR, HRR, and basic anthropometrics, the model achieves state-of-the-art predictive performance. A key finding of this work is the importance of incorporating anthropometric variables for precise VO_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 VO_2 monitoring beyond laboratory settings.

For example, integration with a wearable spirometry like MiniSpir to gather VE, VT and BF alongside smartwatch collection of HR and HRR parameters could enable minimally-burdensome and affordable predictive testing in field contexts. This has implications for increasing access to VO_2 profiling, and enabling investigation of metabolic responses under real-world conditions rather than confined laboratory protocols.

While prior work has predominantly focused on prediction performance, many of these existing methods can be characterized as "black-box" models that provide little insight into the relationships learned. In contrast, the current study incorporated a variable selection mechanism allowing the model to explicitly determine the relative predictive importance of different features. Such interpretability is valuable, as it provides useful insights for researchers in the domain. For example, through analyzing feature attention weights and importance metrics, we found variables like VT to be more influential in predictions than other features like BF and HRR. This finding aligns with theoretical understandings of the key physiological determinants of cardiorespiratory responses.

Being able to reveal predictive relationships in a transparent, quantifiable manner can guide future data collection efforts to target the most salient metrics.

Moreover, interpretability facilitates iterative model development. The model's predictions can be scrutinized against current physiological knowledge to identify potential areas for refinement. Optimization of model architecture, hyperparameters or learning objectives can then be informed by domain expertise. Over time, this process of post-hoc analysis and targeted improvement promises to yield increasingly accurate and well-calibrated forecasts.

Overall, our results demonstrate the potential for interpretable AI to leverage wearable-accessible indicators as a pathway to advancing non-exercise VO_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 VO_2 analyses in medical and health domains.

REFERENCES

[1] T. Tamura, "Wearable oxygen uptake and energy expenditure monitors," *Physiological Measurement*, vol. 40, no. 8, p. 08TR01, 2019.

[2] T. Ebihara, K. Shimizu, M. Ojima, Y. Nakamura, Y. Mitsuyama, M. Ohnishi, H. Ogura, and T. Shimazu, "Energy expenditure and oxygen uptake kinetics in critically ill elderly patients," *Journal of Parenteral and Enteral Nutrition*, vol. 46, no. 1, pp. 75–82, 2022.

[3] R. Malhotra, K. Bakken, E. D'Elia, and G. D. Lewis, "Cardiopulmonary exercise testing in heart failure," *JACC: Heart Failure*, vol. 4, no. 8, pp. 607–616, 2016.

[4] S. Caravita, I. Tanini, L. Crotti, C. Baratto, G. Parati, F. Fattirolli, I. Olivotto, and F. Cecchi, "Impaired cardiopulmonary test performance as a marker of early functional impairment in patients with anderson-fabry disease," *Journal of Cardiovascular Medicine and Cardiology*, pp. 069–071, 11 2021.

[5] M. Bonato, S. Rampichini, M. Ferrara, S. Benedini, P. Sbriccoli, G. Merati, E. Franchini, A. 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*, vol. 55, no. 11, pp. 1277–1284, 2015.

[6] S. E. Crouter, S. R. LaMunion, P. R. Hibbing, A. S. Kaplan, and D. R. Bassett Jr, "Accuracy of the cosmed k5 portable calorimeter," *PLoS One*, vol. 14, no. 12, p. e0226290, 2019.

[7] M. Petelczyc, M. Kotlewski, S. Bruhn, and M. Weippert, "Maximal oxygen uptake prediction from submaximal bicycle ergometry using a differential model," *Scientific Reports*, vol. 13, no. 1, p. 11289, 2023.

[8] K. Przednowek, Z. Barabasz, M. Zadarko-Domaradzka, K. H. Przednowek, E. Nizioł-Babiarz, M. Huzarski, K. Sibiga, B. Dziadek, and E. Zadarko, "Predictive modeling of vo2max based on 20 m shuttle run test for young healthy people," *Applied Sciences*, vol. 8, no. 11, p. 2213, 2018.

[9] F. Abut, M. F. Akay, and J. 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*, vol. 27, no. 5, pp. 3648–3664, 2019.

[10] ——, "Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection," *Computers in biology and medicine*, vol. 79, pp. 182–192, 2016.

[11] B. Carrier, A. Creer, L. R. Williams, T. M. Holmes, B. D. Jolley, S. Dahl, E. Weber, and T. Standifird, "Validation of garmin fenix 3 hr fitness tracker biomechanics and metabolics (vo2max)," *Journal for the Measurement of Physical Behaviour*, vol. 3, no. 4, pp. 331–337, 2020.

[12] C. Spaccarotella, A. Polimeni, C. Mancuso, G. Pelaia, G. Esposito, and C. 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*, vol. 11, no. 6, p. 1467, 2022.

[13] Z. I. Attia, D. M. Harmon, J. Dugan, L. Manka, F. Lopez-Jimenez, A. Lerman, K. C. Siontis, P. A. Noseworthy, X. Yao, E. W. Klavetter *et al.*, "Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction," *Nature medicine*, vol. 28, no. 12, pp. 2497–2503, 2022.

[14] J. Montes, J. C. Young, R. Tandy, and J. W. Navalta, "Reliability and validation of the hexoskin wearable bio-collection device during walking conditions," *International journal of exercise science*, vol. 11, no. 7, p. 806, 2018.

[15] A. Zignoli, A. Fornasiero, M. Ragni, B. Pellegrini, F. Schena, F. Biral, and P. 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*, vol. 15, no. 3, p. e0229466, 2020.

[16] R. Amelard, E. T. Hedge, and R. L. Hughson, "Temporal convolutional networks predict dynamic oxygen uptake response from wearable sensors across exercise intensities," *NPJ digital medicine*, vol. 4, no. 1, p. 156, 2021.

[17] M. Altini, J. Penders, and O. Amft, "Estimating oxygen uptake during nonsteady-state activities and transitions using wearable sensors," *IEEE journal of biomedical and health informatics*, vol. 20, no. 2, pp. 469–475, 2015.

[18] M. M. H. Shandhi, W. H. Bartlett, J. A. Heller, M. Etemadi, A. Young, T. Plötz, and O. 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*, vol. 25, no. 3, pp. 634–646, 2020.

[19] A. Ashfaq, N. Cronin, and P. Müller, "Recent advances in machine learning for maximal oxygen uptake (vo2 max) prediction: A review," *Informatics in Medicine Unlocked*, vol. 28, p. 100863, 2022.

[20] S. W. Su, L. Wang, B. G. Celler, and A. V. Savkin, "Oxygen uptake estimation in humans during exercise using a hammerstein model," *Annals of biomedical engineering*, vol. 35, pp. 1898–1906, 2007.

[21] A. J. Cook, B. Ng, G. D. Gargiulo, D. Hindmarsh, M. Pitney, T. Lehmann, and T. J. Hamilton, "Instantaneous vo2 from a wearable device," *Medical Engineering & Physics*, vol. 52, pp. 41–48, 2018.

[22] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[23] J. L. Fleg and E. G. Lakatta, "Role of muscle loss in the age-associated reduction in vo2 max," *Journal of applied physiology*, vol. 65, no. 3, pp. 1147–1151, 1988.

[24] 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*, vol. 56, no. 6, pp. 725–736, 1987.