Utilising machine learning to optimise the prediction of physiological performance parameters

Jorge Oca Ramirez
MSc Data Science with Artificial Intelligence
University of Exeter
Exeter, UK
jo444@exeter.ac.uk

***Abstract*—This study aims to develop machine learning models capable of accurately predicting Critical Power (CP) and the finite energy reserve (W') with a reduced data window. The results of this study reveal significant advances in the prediction of CP using participant descriptors, power, VO2, and peak data sets, achieving a MAPE of 6.34% for CP and 15.79% for W' using the first 30 seconds of the power data set. The random forest model achieved a 45% improvement in CP prediction accuracy compared to the baseline model. The findings of this research not only shed light on the physiological relationships but also offer practical implications for athletes and researchers. As the methodology evolves and data collection techniques improve, further advancements in predicting these performance parameters can be expected, potentially revolutionising training and performance strategies in the realm of sports and exercise physiology.**

***Index Terms*—Critical Power (CP), Fixed energetic reserve (W'), Performance, Machine learning models.**

I. Introduction

Athletes, coaches, and sports scientists share a common goal: optimising athletic performance. A crucial aspect of this aim lies in understanding each athlete's capability and performance to develop effective pacing and tactical strategies, ultimately maximising the likelihood of success in competitions. The relationship between power (or speed) and the amount of time exercise can be continued at this power is essential for this understanding.

In sports science, the concept of critical power (CP) defines the power threshold beyond which an athlete's energy reserve, denoted by W', becomes depleted. This fundamental metric plays a pivotal role in measuring an athlete's endurance and performance potential. Traditionally, determining CP and W' has relied on exhaustive tests, which may negatively impact an athlete's performance and recovery.

To overcome this limitation and revolutionise the assessment of athletes' capabilities, our project aims to develop machine learning models capable of accurately predicting CP and W' in the context of cycling performance. By doing so, we seek to substantially reduce the duration of testing procedures, offering a less intrusive and more practical approach. This breakthrough would not only benefit athletes by minimising fatigue and recovery time but also hold considerable significance for the field of cycling and sports science.

Currently, CP and W' assessments rely on 3-minute exhaustive tests. Our ambition is to shorten this testing duration by at least 50% while maintaining the accuracy of predictions. Achieving this goal is not without its challenges; predicting important performance measures from limited data presents a special challenge for machine learning.

To address this challenge, we will explore various state-of-the-art machine learning algorithms, including but not limited to Linear, Decision Trees, and Random Forest. By comparing the performance of these algorithms, we seek to identify the most efficient and accurate approach to predicting CP and W' with a reduced data window.

The success of our machine learning models has the potential to reach widespread attention, gaining interest from sports organisations, including British Cycling, sports teams, and industrial partners. Our project could lay the foundations for a publication in a high-impact journal in the field of sports science, as well as attract larger grant applications to further explore and deepen our understanding of the intricate relationship between power, duration, and athletic performance in cycling.

Accurately predicting CP and W' with a reduced data window holds the promise of revolutionising training methods, enhancing athlete performance, and informing coaching strategies, therefore advancing the field of sports science and positively impacting sports organisations and industrial partners.

II. Background

The pioneering work of A.V. Hill in 1925 marked the beginning of investigations between power and duration in athletic events. Hill's groundbreaking research focused on figuring out the fundamental relationship between power output and the duration of various physical activities. Through his extensive exploration, Hill introduced the critical power concept, a key concept that represents the uppermost limit of sustainable power production during exercise [1].

The critical power (CP) is characterised by the slope of the power-duration relationship, a key factor in predicting exercise performance. Hill's research led to the identification of a finite energy reservoir, denoted as W', which represents the amount of energy that can be expended above CP before exhaustion sets in. The relationship between CP and W' is known as W'BAL, and it has become an important factor in determining

an individual's ability to keep pushing and ultimately excel in sports endeavours.

Following A.V. Hill's work, extensive research has been conducted to explore the power-duration relationship, defined by critical power (CP) and the finite energy reserve (W'). In 1965, Monod and Scherrer introduced a significant model, presenting a linear relationship between work done and time to exhaustion, with CP as the slope and W' as the y-intercept, as illustrated in Figure 1 [2].

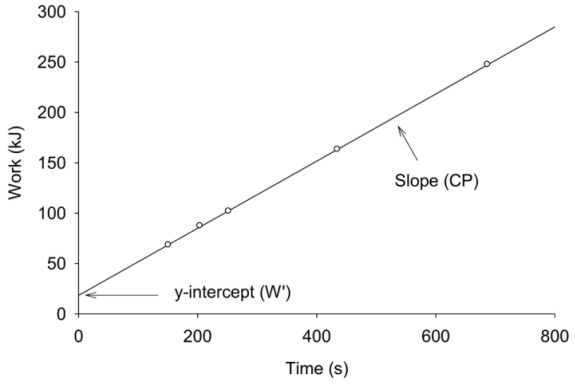

Fig. 1. Linear relationship between work done and time to exhaustion.

In 1991, a study conducted by David G. Jenkins and Brian M. Quigley [3] explored the correlation between the y-intercept of the critical power curve and various measures of anaerobic work capacity. The findings of their research indicated that the y-intercept of the critical power function serves as a reliable indicator of anaerobic work capacity. It signifies an individual's ability to handle both endurance activities and high-intensity work. Understanding this relationship between anaerobic capacity and the critical power function holds significance in evaluating exercise capacity and performance during physically demanding tasks.

Jones et al. [4] proposed an alternative approach to measuring W', as before it was said that CP is overestimated, although this is not the case [5]. They suggested a curvilinear relationship between W' and CP, contrary to the previous linear assumption. This finding implies that W' does not decrease at a constant rate but exhibits a more complex pattern during exercise as shown in Figure 2. As a result, a balance between depleting W' before finishing an event and maintaining optimal performance is essential to achieve success [6]. Furthermore, Jones et al. [7] found that the use of mobile power meters for determining CP and W' can be used for accurately predict performance in 16.1 km cycling events.

Several factors have been identified to influence exercise tolerance, notably including VO2 max and aerobic capacity, which play a role in sustaining CP levels during exercise. The research has demonstrated that W' diminishes with exercise, and this reduction occurs at a higher rate when exercise is performed above CP, reflecting the finite nature of this energy reserve [8].

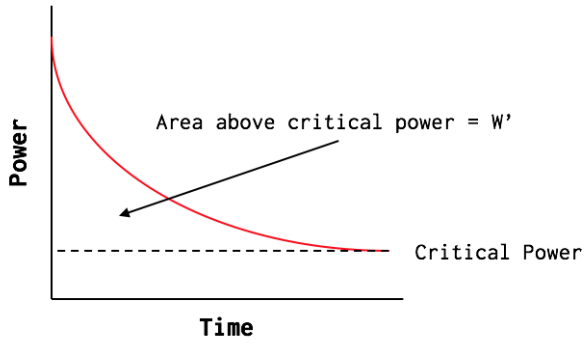

Fig. 2. Current representation of CP and W'.

Vanhatalo et al. introduced a novel approach to accurately predict CP from end power, utilising a 3-minute all-out exhaustion test. This test has shown promising results, providing accurate CP estimations with minimal noise and strong research outcomes [9]. Additionally, the test has been applied successfully in predicting time trials in cycling road races [10]. Understanding that CP and W' are not constant during exercise but vary with intensity domains has highlighted the importance of real-term monitoring to assess changes in these parameters [11]. To demonstrate that this was successful, participants were required to maintain their VO2 profiles while working under this estimated CP threshold for 30 minutes. They were working below CP during the entire duration of the test, so this demonstrated to be a successful finding. The 3-minute test can be used to determine the VO2 Max, which is the peak oxygen intake and end test power, according to Burnley et al. [12]. Furthermore, the research by Vanhatalo et al. [13] revealed that W' is not strongly correlated with specific muscle fibre types, implying a complex interrelationship of various factors in determining exercise capacity. Another study by Clark et al. [14] found that CP does not decrease if the athlete keeps ingesting a sufficient amount of carbohydrates during exercise, although this does not apply for W' as they found that it kept decreasing.

Since A.V. Hill's groundbreaking work, the study of the power-duration relationship, including CP and W', has advanced significantly. The linear model initially proposed by Monod and Scherrer has been challenged, with the dynamic nature of CP and W' becoming apparent. The advancements in predictive techniques, such as the 3-minute all-out exhaustion test, have improved our ability to estimate CP accurately. Additionally, the understanding of the curvilinear relationship of W' and the various factors influencing exercise tolerance has provided valuable insights. The potential for improving training routines and maximising athletic performance in physically demanding tasks is very high.

In a recent study conducted by Ebel et al. [15], the focus was on accurately predicting CP and W' by utilising the shortest duration of the 3-minute all-out tests developed by Vanhatalo et al. Employing PCA (Principal Component Analysis) and

a Ridge regression model, they achieved promising results. They proved that it is possible to predict CP and W' with an approximate error ranging from 13% to 18%, while achieving an 8% to 10% error for CP prediction without relying on the 3-minute all-out tests. These findings highlight the potential of using efficient and effective predictive modelling techniques in the field of exercise physiology, opening up new possibilities for optimising training and performance assessment.

Due to the limited literature on predicting CP and W' using machine learning, our focus in this report is on developing an innovative model rather than relying on existing methods.

III. Aims & Objectives

The primary aim of this project is to design and develop a machine learning model that can effectively identify the minimum data necessary to accurately determine critical power (CP) and the finite energy reserve (W'). The intention is to assess the necessity of the current 3-minute all-out test used for determining CP and W' and, if possible, reduce its duration without compromising accuracy.

- Create a robust machine learning model capable of learning the complex relationships between input features and CP/W', enabling accurate predictions.
- Compare the results obtained by the machine learning model with those from the traditional 3-minute all-out test to determine if the model can predict CP and W' with reduced testing time.
- If time permits, investigate the feasibility of implementing the machine learning model for real-time performance monitoring during athletic events. This will enable the prediction of performance and optimisation of strategies during competitions.

By accomplishing these aims and objectives, this project seeks to advance predictive modelling in exercise physiology and fill the research gap through experimentation and evaluation. It offers a more efficient and accurate method for assessing exercise performance and aiding athletes in reaching their full potential.

IV. Experiment Design & Methods

A. data set & Tools

We will be using data collected by the University of Exeter Sport and Health Sciences department. This data is split into four data sets. The first one consists of biometric information (sex, age, body mass (bm), and height) and features from a ramp test (VO2 peak, power peak, get(L.min-1), and get(W)). Each participant must perform one ramp test before doing the 3-minute all-out test so that the resistance can be adjusted accordingly. The other three data sets contain the values for the 3-minute all-out test data (power, cadence, and oxygen consumption / VO2). The values for these last three data sets were recorded every second, meaning there are 180 columns in each data set(one column for each second of time of the test).

There are a total of 463 tests recorded and 126 participants (some participants have done more than one 3-minute all-out test). The participant descriptors data set described above is free from any personally identifiable information, such as names, addresses, phone numbers, or dates of birth. The absence of this information makes it impossible to link the data back to the individuals involved to ensure that the privacy of the individuals is maintained. In conclusion, the data set provided by the University of Exeter Sport and Health Sciences department is an ideal resource for our project, as it presents minimal to no risk to the participants and this research.

B. CP & W' Calculations

Before starting the analysis and prediction, we need to calculate the CP and W' for each test. We calculate CP by taking the mean power output during the last 30 seconds of the 3-minute all-out test. W' is the work done above CP, and for that, we need to first determine CP, then determine the amount of work performed if the whole test was performed at CP, and finally subtract the work done at CP from the total work performed during the 3-min all-out test.

C. Data Pre-processing

In order to work with the data, we will initiate the process by importing the Excel files into Jupyter Notebooks. To facilitate this task, DataFrames will be employed. The initial phase of the project entails the preprocessing of the data sets. This encompassing stage involves the following key steps:

- Data cleaning: Ensuring a consistent data set by validating uniform formatting and adherence to standardised units of measurement. This step also involves identifying and addressing missing values, outliers, and duplicates.
- Data transformation: Converting the raw data into a format suitable for comprehensive analysis. Techniques such as standardisation will be applied to normalise the data, and any nominal data will be transformed into ordinal data as needed.
- Data integration: Combining data from different data sets to create one cohesive, unified data set that allows for more comprehensive insights.

When visualising the participant descriptors and VO2 DataFrames, we observed that values from the Get(L.min-1) column (participant descriptors data set) and some other VO2 values (VO2 data set) did not follow a uniform unit of measurement. To address this discrepancy, adjustments were made to ensure uniformity in tenths (L.min-1 instead of ml.min-1). The participant descriptors Dataframe had some null values which we converted to zeros for future reference.

The power, cadence, and VO2 DataFrames also presented instances of null values. The approach to solving this was to fill the empty values with the average of the preceding and succeeding non-empty values.

Within the VO2 data set, a discrepancy in data recording intervals was discovered, with some rows reflecting measurements every 10 seconds and others every second. We used linear interpolation to balance this irregularity and guarantee data availability for every second.

These stages collectively establish a solid foundation for the subsequent utilisation of machine learning techniques to optimise the prediction of physiological performance parameters.

D. Data Visualisation

In order to gain insights into the nature of relationships, we plan to create plots that illustrate the connection between specific features and CP or W'. We started with plotting a correlation matrix which can be seen in Figure 3. This matrix focuses on the participant descriptors in relation to CP and W', allowing us to grasp initial patterns and associations.

	CP	w
sex	0.216391	0.272762
age(y)	0.127577	-0.059190
bm(kg)	0.086771	0.126471
height(m)	0.103873	-0.010341
vo2_peak(L.min-1)	0.629745	0.248819
peak_power(W)	0.652372	0.192698
get(L.min-1)	0.023528	-0.100415
get(W)	0.451978	0.109182

Fig. 3. Correlation matrix of the participant descriptors with CP and W'.

The insights drawn from Figure 3 reveal strong correlations between CP and both peak power and VO2 peak. Additionally, a moderate correlation is observed between CP and the GET(W) variable, all of which are derived from the initial ramp test. On the other hand, there is a noticeably smaller link between participant descriptions and W'.

Furthermore, we created correlation matrices to assess the relationships between the power, cadence, and VO2 DataFrames with CP and W'. Notably, the power correlation matrix reveals insightful trends: CP exhibits a moderate correlation ranging from 0.4 to 0.5 during the initial 30 seconds, followed by a linearly increasing strong correlation from 0.5 to 0.99, extending until the 180th second of the 3-minute all-out test. In contrast, W' displays a strong correlation, ranging from 0.55 to 0.85, within the initial 20 seconds, followed by a decline to 0.2 at the 180-second mark of the test.

Surprisingly, there are no observable relationships between CP or W' revealed by the cadence correlation matrix. Finally, analysing the VO2 correlation matrix, we note a moderate correlation between VO2 and CP spanning seconds 20 to 30 (0.21-0.48), followed by a progressively strong correlation between seconds 30 and 180 (0.51-0.73). W', on the other hand, shows no obvious correlation with the VO2 data set.

E. Feature Engineering

Enhancing the prediction capability and interpretability of machine learning models relies heavily on feature engineering. This section explores the strategic practice of feature engineering in the context of our effort to optimise the prediction of physiological performance parameters using machine learning techniques.

We aim to turn raw data into meaningful representations that capture the underlying patterns and relationships by carefully creating and selecting important features from our data set. By carefully extracting, transforming, and combining variables, we might potentially uncover hidden insights and provide our models with the tools they need to better understand the complexity involved in predicting physiological performance.

The current features present in the DataFrames do not account for the evolving relationship between consecutive seconds. In contrast, the significant association between CP and W' captures the temporal fluctuations that span each individual second.

Considering the 3-minute test's objective of identifying the point where the curve's gradient levels off (CP), we can explore the incorporation of gradients between each data point. Furthermore, W' can also be thought of as the integral under the power curve, suggesting a potential link between the derivative and W'.

As highlighted during the data visualisation phase, the peaks hold notable importance when predicting these features. To address this, we can compute the cumulative sum, allowing the power values to progressively accumulate over time while still maintaining the primary peaks. This approach guarantees we are still able to calculate the gradients at each time point as it protects the important peaks.

During the data visualisation phase, we also discovered that the power and VO2 max peaks were highly correlated with CP, therefore we can calculate the peaks for power and VO2 between time 0 and 't', 't' being the maximum amount of seconds of the 3-minute test used during the prediction phase.

F. Train-Test Split

In order to partition our data set for training and testing, a 70-30 split ratio was applied, ensuring a balanced allocation of data.

The process commences by shuffling the input Dataframe to introduce randomness, an essential step to mitigate potential biases. Subsequently, participants are identified along with the count of tests they have undergone. A systematic approach is then employed to allocate individuals into the train and test sets while preserving the grouping of tests belonging to the same participant.

This rigorous allocation produces the final train and test DataFrames, which contain rows relating to the designated participants. These DataFrames are further segregated into feature matrices (X_train, X_test) and target vectors (y_train, y_test).

Standardisation, a crucial preprocessing step to enhance model convergence and performance, is then executed. The feature matrices (X_train, X_test) undergo scaling using the StandardScaler function from the SkLearn python library to ensure uniformity in scale across the features.

In essence, this technique maintains the fundamental links between tests taken by the same individual within each split while adhering to a well-structured 70-30 train-test split.

G. Models

In this section, we present an in-depth analysis of the various models we have used to try to improve the accuracy of our predictions of physiological performance parameters. Each model was carefully selected to serve a certain function while offering a variety of viewpoints and performance standards. The selection process considered the need for interpretability, predictive accuracy, and the ability to capture underlying complexities. The models introduced here are a baseline model for comparison, and advanced techniques like linear regression, decision trees, and random forests.

The baseline model serves as a benchmark against which the performance of more complex models is measured. By creating a baseline, we obtain an understanding of the lowest level of predictability that is achievable, assisting in the assessment of the improvements achieved by more complex models.

The linear regressor was chosen for its simplicity and interpretability. Our objective of understanding the linear correlations between features and the target variable is consistent with its basic premise. While linear regression may not capture complex nonlinear patterns, it offers a precise interpretation of the effects of features and makes it easier to understand how various physiological parameters relate to one another.

Decision trees provide a simple way to visualise data relationships. They are capable of adjusting to nonlinear patterns and interactions within the data. Decision trees allow us to visualise the decision-making process and also provide an additional layer of understanding and transparency in our prediction models.

Finally, the random forest model was selected to maximise the potential of various distinct models working together. By combining the predictions of multiple decision trees, random forest mitigates the overfitting tendencies of individual trees and enhances generalisation to new data. This model's ability to capture complex relationships, manage overfitting, and maintain interpretability fits with the main goal we have in mind.

To summarise, the diverse selection of models employed in our analysis ensures a robust exploration of predictive capabilities. Each model's unique strengths contribute to a comprehensive understanding of the complex hidden physiological performance parameters relationships. These models' comparison and evaluation will help us in looking for the best prediction by showing their respective strengths.

H. Error metrics & Cross Validation

To understand the prediction capabilities of the models described above and being able to compare them to each other, we will be using error metrics like the Mean Squared Error (MSE), R-Squared(R2), and Mean Absolute Percentage Error (MAPE) [16].

The MSE metric represents the average of the squared differences between the predicted values and the actual values. The MSE penalises large errors more heavily, meaning we should be careful with having outliers in the data.

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

- n is the number of data points
- y_i is the actual value for the i$^{\text{th}}$ data point
- $\hat{y}_i$ is the predicted value for the i$^{\text{th}}$ data point.

The R2 metric measures the proportion of the variance in the dependent variable (targets) that is explained by the independent variables (features) in the model. It ranges from 0 to 1, with higher values indicating a better fit.

$$R^2 = 1 - \frac{SS_{residual}}{SS_total}$$

- $SS_{residual}$ is the number of data points
- SS_{total} is the actual value for the i$^{\text{th}}$ data point

Finally, the MAPE calculates the average percentage difference between the predicted values and the actual values. It gives a direct and interpretable measure of prediction accuracy in terms of percentage error, which can be easily understood.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

- n is the number of data points
- y_i is the actual value for the ith data point
- $\hat{y}_i$ is the predicted value for the ith data point.

When looking at the model results we will be using these 3 metrics to help us determine the best model for predicting the performance parameters. We will also be using k-fold Cross-Validation to guard against overfitting and to have a reliable estimation of model performance on unseen data. If k is 5, the model will be iteratively trained on four folds while validated on the fifth. This procedure will be repeated five times, making sure each fold serves as the validation set once.

I. Model Hyperparameters

The adjustment of hyperparameters is crucial in our effort to improve model performance [17]. Hyperparameters are parameters that are not learned from the data but are set before the learning process begins. In this context, our approach to hyperparameter tuning contributes significantly to the optimisation of predictive accuracy.

When it comes to decision trees, our attention is drawn to the modification of two key hyperparameters [18]:

- Max Depth: This hyperparameter controls the maximum allowable depth of the decision tree. By adjusting it, we influence the extent to which the tree can capture complex relationships within the data, preventing overfitting.
- Min Samples Leaf: The min samples leaf hyperparameter defines the minimum number of data samples required to

form a leaf node. Adjusting this parameter allows us to control the granularity of the decision tree, potentially impeding overfitting tendencies.

In the case of random forests, a thorough exploration of the following hyperparameters takes place [19]:

- Number of Estimators: This parameter controls the number of individual decision trees joined within the random forest ensemble. Balancing this value affects the ensemble's ability to generalise across the data set.
- Max Depth: Similar to decision trees, the max depth hyperparameter for random forest manages the maximum depth of each decision tree. Carefully adjusting this value can influence the ensemble's performance and its resistance to overfitting.
- Min Samples Leaf: This hyperparameter, as with decision trees, contributes to the granularity of leaf nodes in individual trees. By tuning it, we impact the overall forest's adaptability to the data.

In our search for the best hyperparameter settings, we employ a robust approach: 5-fold cross-validation. This method prevents overfitting while ensuring a comprehensive evaluation of various hyperparameter setups. By partitioning the data set into five distinct subsets, we iteratively train the model on four folds and validate it on the remaining fold. This process is repeated five times, ensuring that each fold serves as the validation set at least once. The resulting evaluations provide a reliable estimation of model performance on unseen data, offering insights into the generalisation capabilities of different hyperparameter settings.

In summary, the adjustment of hyperparameters, accompanied by cross-validation, constitutes a crucial part of our methods. This strategy gives us the ability to utilise decision trees and random forests to their fullest extent, enhancing predictive accuracy and promoting model generalisation.

V. Results

In this section, we present the results achieved through the utilisation of the diverse models that have been developed. The next section will focus on explaining these results and each model's performance. Our evaluation protocol ranks these models primarily based on the MAPE score, followed by R2, and lastly, MSE. This ranking framework is established with a strong focus on predictive accuracy, followed by the model's ability to fit the data. We have also identified a time window of 60-90 seconds as the cutoff threshold for the predictions as the aim of this study is to reduce the duration of the 3-minute all-out test.

Starting with the baseline model, the best result recorded had an MSE of 3920.43, an R2 of -0.56, and a MAPE of 18.20% when predicting CP, and 28.43, -0.11, and 33.35% respectively for predicting W' using the power data set. Looking at all the different baseline models created, the MAPE for CP ranges from 18.20% to 25.46%, and from 33.35% to 45.88% for W'.

The insights from the baseline model will serve as a foundation, helping in the explanation and comparison of the following models' results. By establishing a reference point, the baseline model facilitates a comprehensive understanding of the improvements achieved. We know that a successful model for CP and W' should be below 18% and 33% respectively.

The best-performing linear model obtained an MSE of 963.56, R2 of 0.61, and MAPE of 20.53 for CP using the participant descriptors data set, and 28.44, -0.06, and 34.84% for W' using the first 10 seconds of the VO2 data set.

The decision tree algorithm obtained the best result for CP using participant descriptors and the first 30 seconds of the power data set, obtaining 1288.03 for MSE, 0.47 for R2 and 10. 43% for MAPE. For W', the best result was obtained using the first 20 seconds of the power data set, with the MSE, R2, and MAPE scores of 11.97, 0.54, and 20.98%.

Finally, the random forest model obtained the best prediction overall for CP using participant descriptors, and the first 60 seconds of the peaks and derivatives. This model obtained an MSE of 786.02, an R2 of 0.76, and a MAPE of 8.46%. For W', we also obtained the best overall results using the participant descriptors and the first 30 seconds of the power data set. We obtained an MSE of 5.78, an R2 of 0.77, and a MAPE of 15.79%.

Figures 4 and 5 present the MAPE scores obtained by the best-performing models using each algorithm (baseline, linear, decision tree, random forest), throughout the training time from 0 to t. All these best-performing models have been presented above.

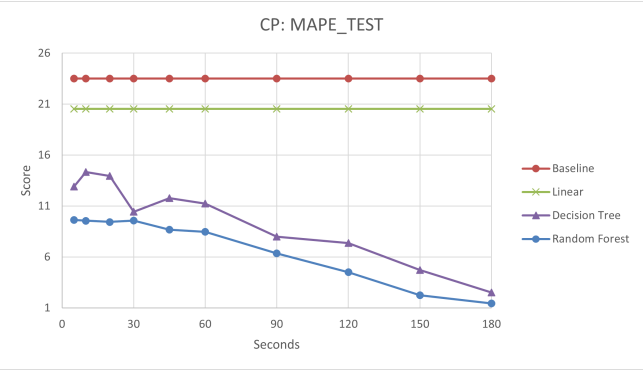

Fig. 4. Best models: MAPE when predicting CP.

Furthermore, in contrast, Figures 6 and 7 display the MAPE scores associated with the worst-performing models using each algorithm (Baseline, Linear, Decision Tree, Random Forest) during the training interval from time 0 to 't'. The worst-performing models when predicting CP include:

- The baseline model, exhibiting a MAPE of 25.46% when utilising participant descriptors, power, and cadence data.
- The linear model and decision tree models, each with an average MAPE of 32.3% and 24.8%, when utilising participant descriptors, power, cadence, and VO2 data, and the cadence data set, respectively.

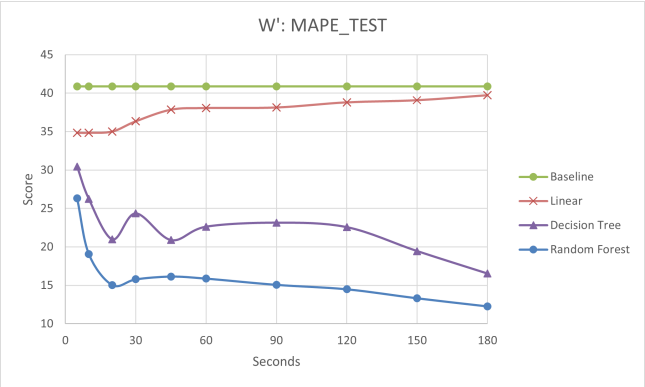

Fig. 5. Best models: MAPE when predicting W'.

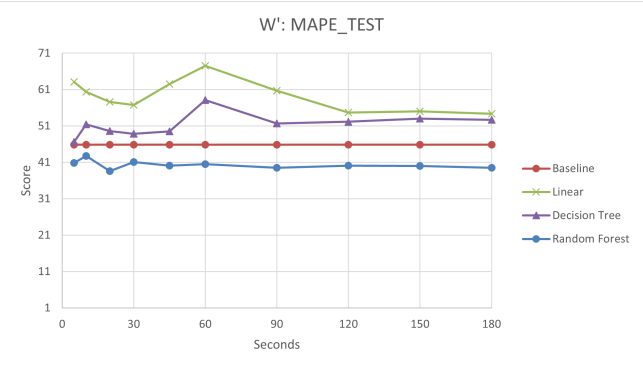

Fig. 7. Worst models: MAPE when predicting W'.

- The random forest model, which showed a marginally improved performance compared to the baseline, with a MAPE of 20.15% when trained on the cadence data set.

Similar observations are mirrored in W' prediction:

- The baseline model returns a MAPE of 45.88% employing participant descriptors, power, and cadence data.
- Both the linear and decision tree models generate average MAPEs of approximately 59% and 50%, using participant descriptors, power, cadence, and VO2 data, and the cadence data set, respectively.
- The random forest model is the most successful among these, registering an approximate MAPE of 40% when using the cadence data set alone.

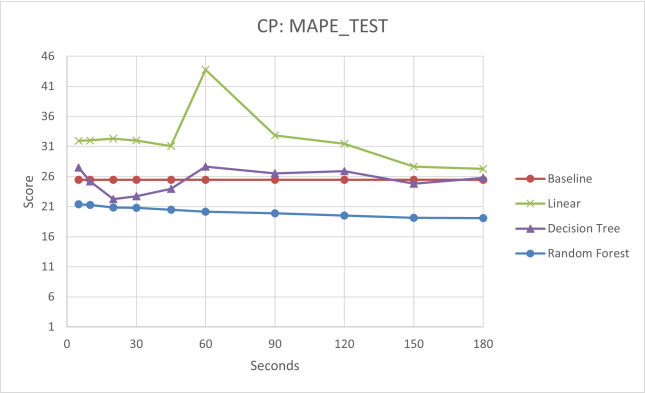

Fig. 6. Worst models: MAPE when predicting CP.

These findings highlight the complexities involved in predicting CP and W', illustrating the strengths and limitations of various models while guiding the evolution of our predictive approach.

VI. Discussion

In this section, we provide a thorough analysis of the key findings of our investigation, comparing them with existing related works to reveal understandings, similarities, differences, and implications. This discussion focusses on limitations and methodological flaws while exploring the cross-disciplinary importance and potential commercial applications of our predictive models to clarify the broader implications of the results of our investigation.

Our study seeks to advance the prediction of physiological performance parameters, specifically Critical Power (CP) and W', whilst reducing the duration of the 3-minute all-out tests by applying machine learning models. In the context of related work, as we mentioned before, there are not many studies which have studied the prediction of CP and W' whilst reducing the duration of the tests. Comparing our results with a study by Ebel et al. [15], our results demonstrate several noteworthy parallels and distinctions.

To start with, we were able to predict CP just with the participant descriptors data set with a MAPE of 9.98% using a random forest model. This prediction of CP shows a 45% improvement in accuracy compared to the initial baseline model developed. We can attribute these results to the fact that the variables of peak power, peak VO2 and power (GET (W)) obtained from the ramp test were highly correlated with CP as we saw in the correlation matrix 3.

The best-performing model when predicting CP, using random forest, obtained an 8.46% MAPE when using the participant descriptors, and the first 60 seconds of the duration of the 3-minute all-out test for calculating the peaks and derivatives. This means a 15.23% increase from the random forest model only using the participant descriptors, and a 53.52% increase in accuracy when compared with the baseline model. It is possible to obtain a more accurate prediction. To do this, we can increase the duration of the test to the first 90 seconds, and we will be able to get a prediction with an error of 6.34% for CP, using a random forest model, using the participant descriptors data set, and the first 90 seconds of the test to calculate the peaks and derivatives. As we can see in Figure 4, we start getting a major improvement in the MAPE metric, after the first 45-60 seconds of the duration of the test. This is explained by the fact that the correlation between the power data set and CP starts linearly increasing after the first 30 seconds of the 3-minute all-out tests. We observe something similar with the VO2 data set, where the correlation starts increasing after the first 30 seconds of the test.

Moving on to W', the baseline model obtained an accuracy of 33.35%. When using just the participant descriptors data set and a random forest model, we obtained a worst performance if we compare it with the baseline model using all the data sets. This can be explained as we did not observe any strong correlations between the participant descriptors and W', as seen in Figure 3.

The best result, when predicting W', was obtained using a random forest model, along the participant descriptors data set and the first 30 seconds of the power data set. The model obtained a 15. 79% error, which can be explained as we saw a strong correlation between W' and the first 20 seconds of the power data set, but no strong correlations with the participant descriptors.

It appears that CP is easier to predict with less data, in comparison to W', where it is difficult to find features that correlate with W'. We were able to find correlations between CP and the participant descriptors, power, VO2, and peaks data sets, but just with the power data set for W'.

Future research should focus on finding and creating new features based on the data we currently have to discover more relationships with CP, but especially with W' as we were not able to do so. Also, it should be important to make sure the data collection is performed correctly so all the values needed are stored in a consistent manner as we found some tests from participants with missing values for VO2 and cadence. Because of this, we were forced to delete these tests in some cases where it was not possible to fill in the missing values. Another method for improvement could be utilising Principal Component Analysis [20] after creating features for both CP and W', as it can help determine the strongest features for predicting both targets.

The models were trained on an average of 250 tests after removing the ones with missing-null values, although having over 400 at the start. To increase the number of tests used when training the models, we could collect data from mobile power meters [7], which have been shown to be accurate when predicting CP and W'. By using power meters, it would be easier for everyone to do these tests as they would not need to do 5 visits to a lab, just perform the 3-minute all-out tests from their bike.

If the approach above is taken, where data are collected using these mobile power meters, it might be useful to group the data in bins/intervals, as we would probably lose some degree of accuracy, although this needs to be carefully studied. If successful, the potential to get new tests can drastically increase. Another alternative is to implement the models using real-time data to get an idea about the accuracy of predictions during cyclying events.

VII. Conclusion

In conclusion, this study aimed to optimise the prediction of physiological performance parameters, specifically Critical Power (CP) and W', by applying machine learning techniques to reduce the duration of the 3-minute all-out tests. The findings of this study revealed significant advances in the prediction of CP using participant descriptors, power, VO2, and peak data sets. The random forest model achieved a 45% improvement in CP prediction accuracy compared to the baseline model. Furthermore, by extending the duration of the test to the first 90 seconds, the prediction accuracy increased to just a 6.34% error.

In contrast, predicting W' proved to be more challenging, with limited correlations between participant descriptors and W'. The best performing model for W's prediction involved the use of a random forest model and the first 30 seconds of the power data set, achieving a 15.79% error. This disparity highlights the complexity of predicting W' and underscores the need for future research to uncover additional features and relationships.

To enhance future predictive models, it is recommended to explore new data features based on the existing data set, especially for W', and ensure consistent data collection to mitigate missing values. Utilising Principal Component Analysis (PCA) may help identify stronger features for predicting both CP and W'. Furthermore, the potential integration of data from mobile power meters presents an opportunity to expand the data set and make testing more accessible, thus improving the accuracy of the models.

In conclusion, this study contributes to the field of sports science by demonstrating the feasibility of predicting CP and W' using machine learning models while reducing the testing duration. The insights gained from this research not only shed light on the physiological relationships, but also offer practical implications for athletes and researchers. As the methodology evolves and data collection techniques improve, further advancements in predicting these performance parameters can be expected, potentially revolutionising training and performance strategies in the realm of sports and exercise physiology.

VIII. Declarations

Declaration of Originality. I am aware of and understand the University of Exeter's policy on plagiarism and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices.

Declaration of Ethical Concerns. This work does not raise any ethical issues. No human or animal subjects are involved neither has personal data of human subjects been processed. Also no security or safety critical activities have been carried out.

Acknowledgment

I would like to thank Dr. Diogo Pacheco for his help and guidance throughout the project, and Dr. Matthew Black for letting us use their data sets and providing information about them.

References

[1] F.R.S. Prof. A. V. Hill. "The Physiological Basis of Athletic Records." In: *Nature* 116 (2919 Oct. 1925), pp. 544–548.

[2] H. Monod and J. Scherrer. "THE WORK CAPACITY OF A SYNERGIC MUSCULAR GROUP". In: *https://doi.org/10.1080/00140136508930810* 8 (3 1964), pp. 329–338. ISSN: 13665847. DOI: 10.1080/00140136508930810.

[3] DAVID G. JENKINS and BRIAN M. QUIGLEY. "The y-intercept of the critical power function as a measure of anaerobic work capacity". In: *Ergonomics* 34.1 (1991). PMID: 2009846, pp. 13–22. DOI: 10.1080/00140139108967284.

[4] Andrew M. Jones and Anni Vanhatalo. "The 'Critical Power' Concept: Applications to Sports Performance with a Focus on Intermittent High-Intensity Exercise". In: *Sports Medicine* 47 (Mar. 2017), pp. 65–78. ISSN: 11792035. DOI: 10.1007/s40279-017-0688-0.

[5] Andrew M. Jones et al. "The maximal metabolic steady state: redefining the 'gold standard'". In: *Physiological Reports* 7 (10 May 2019).

[6] Brett S. Kirby et al. "Interaction of exercise bioenergetics with pacing behavior predicts track distance running performance". In: *Journal of Applied Physiology* 131 (5 Nov. 2021), pp. 1532–1542. ISSN: 15221601. DOI: 10.1152/japplphysiol.00223.2021.

[7] Andrew M. Jones et al. "The maximal metabolic steady state: redefining the 'gold standard'". In: *Physiological Reports* 7 (10 May 2019).

[8] Mark Burnley and Andrew M. Jones. "Oxygen uptake kinetics as a determinant of sports performance". In: *European Journal of Sport Science* 7 (2 June 2007), pp. 63–79. ISSN: 17461391. DOI: 10.1080/17461390701456148.

[9] Anni Vanhatalo, Jonathan H. Doust, and Mark Burnley. "Determination of critical power using a 3-min all-out cycling test". In: *Medicine and Science in Sports and Exercise* 39 (3 Mar. 2007), pp. 548–555. ISSN: 01959131. DOI: 10.1249/mss.0b013e31802dd3e6.

[10] Matthew I. Black et al. "Critical power derived from a 3-min all-out test predicts 16.1-km road time-trial performance". In: *European Journal of Sport Science* 14 (3 2014), pp. 217–223. ISSN: 15367290. DOI: 10.1080/17461391.2013.810306.

[11] Matthew I Black et al. "Muscle metabolic and neuromuscular determinants of fatigue during cycling in different exercise intensity domains". In: *J Appl Physiol* 122 (2017), pp. 446–459.

[12] Mark Burnley, Jonathan H. Doust, and Anni Vanhatalo. "A 3-min all-out test to determine peak oxygen uptake and the maximal steady state". In: *Medicine and Science in Sports and Exercise* 38 (11 Nov. 2006), pp. 1995–2003.

[13] Anni Vanhatalo et al. "The mechanistic bases of the power–time relationship: muscle metabolic responses and relationships to muscle fibre type". In: *Journal of Physiology* 594 (15 Aug. 2016), pp. 4407–4423. ISSN: 14697793. DOI: 10.1113/JP271879.

[14] Ida E Clark et al. "Dynamics of the power-duration relationship during prolonged endurance exercise and influence of carbohydrate ingestion". In: *J Appl Physiol* 127 (2019), pp. 726–736.

[15] Mark Ebel, Diogo Pacheco, and Matthew Black. *Utilising machine learning to optimise the prediction of physiological performance parameters*. University of Exeter, Aug. 2022.

[16] Alexei Botchkarev. *Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology*.

[17] Li Yang and Abdallah Shami. "On hyperparameter optimization of machine learning algorithms: Theory and practice". In: *Neurocomputing* 415 (Nov. 2020), pp. 295–316.

[18] Rafael G. Mantovani et al. "Hyper-Parameter Tuning of a Decision Tree Induction Algorithm". In: *Proceedings - 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016* (Feb. 2017), pp. 37–42.

[19] Philipp Probst, Marvin N. Wright, and Anne Laure Boulesteix. "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (3 May 2019), e1301.

[20] Takio Kurita. "Principal Component Analysis (PCA)". In: *Computer Vision: A Reference Guide*. Cham: Springer International Publishing, 2019, pp. 1–4.