ORIGINAL ARTICLE

Applying Artificial Intelligence to predict Olympic triathlon performance at Paris 2024 Olympic Games

Pablo García-González . *Physical Performance and Sports Research Center (CIRFD). Pablo de Olavide University. Seville, Spain.*
Luca A. Bianchini. *DBinformation S.P.A. Assago MI, Italy.*
Andrea Fuk. *Department of Movement, Human and Health Sciences. University of Rome "Foro Italico". Rome, Italy.*
Simone Villanova. *Department of Movement, Human and Health Sciences. University of Rome "Foro Italico". Rome, Italy.*
José Antonio González-Jurado. *Physical Performance and Sports Research Center (CIRFD). Pablo de Olavide University. Seville, Spain.*
Maria Francesca Piacentini. *Department of Movement, Human and Health Sciences. University of Rome "Foro Italico". Rome, Italy.*

ABSTRACT

The aim of the present study was to predict triathlon performance at Paris 2024 Olympic Games using conventional statistics and a machine learning-based approach. It was hypothesized that both predictive models would grant sufficiently accurate results. Data were extracted from the API service on the World Triathlon website. A custom Python code was written for the analyses during data collection. Conventional statistics and machine learning analyses were performed by creating a Jupyter Notebook via Google Colab. Data for machine learning were divided into training (80%), and testing (20%). Run time was the best-predicted discipline for males (average difference: -0.21% ± 5.45%), and total time was the best-predicted variable for females (average difference: -5.43% ± 3.81%). For males, the linear regression based on Olympic Games races data was the most accurate technique overall (average relative difference: -3.03% ± 10.79%). For females, TensorFlow achieved the best precision (average relative difference: -1.95% ± 10.43%). Both techniques are reliable approaches for predicting performance in Olympic triathlon. Based on our findings, statistical methods can be effectively employed by coaches to predict individual discipline performance and overall race times. When comparing statistical techniques that considered all data, machine-learning techniques showed better predictions than conventional statistics.

Keywords: Performance analysis, Aerobic endurance, Quantitative data, Multi-sports, Machine learning.

Cite this article as:
García-González, P., Bianchini, L. A., Fuk, A., Villanova, S., González-Jurado, J. A., & Piacentini, M. F. (2026). Applying Artificial Intelligence to predict Olympic triathlon performance at Paris 2024 Olympic Games. *Journal of Human Sport and Exercise, 21*(3), 861-878. https://doi.org/10.55860/py0kyk77

Corresponding author. *Physical Performance and Sports Research Center (CIRFD), Universidad Pablo de Olavide , 41013 Seville, Spain.*
E-mail: pablogartri@gmail.com

Submitted for publication March 25, 2026.
Accepted for publication April 30, 2026.
Published May 07, 2026.
Journal of Human Sport and Exercise. ISSN 1988-5202.

doi: https://doi.org/10.55860/py0kyk77

INTRODUCTION

Triathlon is a sport composed of three disciplines performed non-stop on different distances (short: from supersprint to Olympic distance; long: from middle-distance or Ironman 70.3 to full Ironman or superior). The Olympic distance triathlon involves 1.5 kilometres swim, 40 kilometres bike and 10 kilometres run. Since the invention of triathlon, several researchers have studied the physiological and psycho-social determining factors in Olympic triathlon (O'Toole & Douglas, 1995; Ruiz-Tendero & Salinero Martín, 2012), highlighting the importance of maximal capacity (VO_{2max}), economy of motion (submaximal VO_2), and fractional utilization of maximal capacity (%VO_{2max}) of the triathlete.

Several papers aimed to analyse or predict triathlon performance, either using race data or considering athlete characteristics (García-González & González-Jurado, 2024a, 2024b, 2025b, 2025a; Sousa et al., 2021; Van Schuylenbergh et al., 2004). The most traditional approaches have used anthropometric and physiological variables to model overall and split performance with good accuracy (Van Schuylenbergh et al., 2004). Other authors aimed to predict race outcomes from shorter-distance tests in each discipline, talent-identification batteries, or longitudinal performance trajectories (Cuba-Dorado et al., 2021; Malcata et al., 2014), linking these measures to later competition results and career progression. Recently, performance prediction moved to integrate statistical and machine-learning approaches that give importance to the interaction between disciplines, transitions, and contextual factors such as athlete's pacing strategies and environmental conditions (Olaya-Cuartero et al., 2022; Sousa et al., 2021). This theoretical background shows how triathlon performance prediction can inform individualized training, talent development, and race strategy.

Machine learning (ML) is revolutionizing various sectors, including sports. Artificial Intelligence (AI) is a powerful tool that is reshaping citizen's lives, interactions and environments, transforming profoundly the current habitat (Cath et al., 2018). Several researchers have attempted to apply artificial intelligence to predict certain situations such as musculoskeletal injuries (Bullock et al., 2022). In the context of sport, the impact of AI has been investigated widely (Dindorf et al., 2023; Reis et al., 2024; Sperlich et al., 2023). Several authors have focused on injury prediction in team sports (Claudino et al., 2019; Rossi et al., 2021). Moreover, several mathematical models have also been applied to predict performance in different type of sports (Lim & Song, 2024; Nagovitsyn et al., 2023; Tam & Yao, 2024). In endurance sports, several models have been applied to predict marathon performance, with an accuracy of over 94% (Lerebourg et al., 2022) and to predict daily recovery prediction (Rothschild et al., 2024).

Machine Learning has also been applied to triathlon including different topics: to determine the fastest race course for professional athletes competing in Ironman (Thuany et al., 2024) and Ironman 70.3 (Thuany et al., 2023), to predict running-related injuries in young triathletes (Martínez-Gramage et al., 2020) or to solve the classic problem regarding the most influential discipline in triathlon (García-González et al., 2026; Martínez-Sobrino et al., 2023; Ofoghi et al., 2016; Weiss et al., 2024). In long-distance events, several researchers used tree-based and gradient-boosting regressors trained on large historical datasets of professional Ironman and Ironman 70.3 to rank course locations and environmental conditions by their expected impact on finish times (Knechtle et al., 2025), highlighting the importance of race venue, air and water temperature, and athletes' country of origin when selecting "*fast*" courses. Collectively, these studies show that supervised learning methods are increasingly used to (i) model complex, non-linear relationships between split performances and overall results, (ii) characterize optimal pacing and course selection strategies, and (iii) identify modifiable biomechanical and training-load risk factors, thereby positioning ML/AI as a central

methodological tool for both performance prediction and injury prevention in triathlon (Knechtle et al., 2025; Ofoghi et al., 2016; Thuany et al., 2023; Weiss et al., 2024).

Following an extensive literature review, no prior studies were identified that employed Machine Learning to predict Olympic distance triathlon performance at one specific event.

The aim of the present study is to predict triathlon performance at the Paris 2024 Olympics using conventional statistics and a Machine Learning-based approach. It is hypothesized that both predictive models would grant sufficiently accurate results.

The secondary aim of this study is to compare machine learning methods and conventional statistical technique analyses to evaluate precision and accuracy in predicting triathlon performance outcomes. It is hypothesized that machine learning methods could be more accurate than conventional statistical techniques.

MATERIAL AND METHODS

Participants

Data were extracted from the API service (Application Programming Interface) on the World Triathlon website (https://triathlon.org/) by signing up for the free service. Since the data were in the public domain and available on the Internet, no formal request was made to the Ethics Committee. To perform data collection, custom Python code (Python Software Foundation, Wilmington, DE, USA; ver. 3.12) was written for different operations. The Python codes used are available on the GitHub repository (https://github.com/AEONPHYTON/triathlon_analysis). Once the data were extracted, cleanup operations were performed using Python as the programming language and Jupyter Notebook (Project Jupyter, San Francisco, CA, USA; ver. 7.2) as the working tool (Rule et al., 2019). Two databases were created including Olympic distance races (one for males and one for females) of the elite category. Previous editions of the Olympics were also included.

Files identifying the races by the race number (prog_id and event_id) were downloaded and used to obtain the corresponding results data. The dataset was restricted to Elite and U23 athletes, with U23 retained due to their participation alongside Elite competitors. Standard-distance events were considered, separated by sex. Different variables were generated to track record of each athlete's discipline positions. Athletes lacking data for any discipline or transition, as well as athletes who were disqualified, did not start (DNS), or did not finish (DNF), were removed. Races without complete podium results or with implausible overall times (lower than 1 hour, 28 minutes and 20 seconds and higher than 2 hours and 30 minutes) were also excluded. These procedures granted a heterogeneous but valid dataset including more than 30 years of races, from 6 November 1989 to 24 May 2024.

Procedures

Once the cleaning was completed, two databases (males and females) were obtained for conventional statistical and machine learning analyses. The selected competitions were: Continental Championships, World Championships, World Cup, Continental Cup, Recognised Event, Regional Championships, Recognised Games, Major Games (editions of the Olympics from 2000 to 2021), World Championship Series. The reason behind including different types of competitions was to obtain a large amount of data for analysis. Regarding classical statistical techniques, a linear regression was performed using only the dataset from previous Olympic Games editions. Similarly, another linear regression was conducted using the entire

dataset. Finally, a third-degree polynomial regression was applied to the complete dataset. All the aforementioned analyses were used to estimate performance in the three triathlon disciplines, as well as the overall finish time. Scatter plots representing the distribution of data points were also generated.

To determine the average swimming, cycling, running, and overall times of the podium finishers, a sub-database was created in which only the top 3 finishers were considered for each event and averaged to obtain a unique time for each race; the data were filtered according to the final position of each race. Two databases were created: one with only the Olympic races and the other containing all races (always divided by males and females).

Statistical analyses

Statistical analyses and analyses using machine learning were performed by creating a Jupyter Notebook file (Project Jupyter, San Francisco, CA, USA; ver. 7.2) through the Google Colab service (Google Corp, Mountain View, CA, USA; ver. 2.14). Google Colab provides a faster computation service than regular computers in use to speed up processing. TensorFlow (Google Corp, Mountain View, CA, USA; ver. 2.14) and PyTorch (Meta AI, Menlo Park, CA, USA; ver. 2.3.1) libraries were used for machine learning analyses. For the projection of times to the date of the Paris 2024 Olympics, the last date on which a standard-distance event was held was taken into account, generating an interval of 67 days between that event and the Paris competition.

Data for evaluation through machine learning models was divided into training and testing data with the former used to fit the model and the latter used to assess its predictive performance. The data were split chronologically into 80% training and 20% testing, filtered to fall within the mean ± two standard deviations to reduce the influence of outliers. Machine learning techniques implemented with PyTorch and TensorFlow were employed to analyse data trends with scatter plots generated to visualize data distributions.

RESULTS

In total, the database contained a total of 45,884 and 37,106 male and female triathletes, respectively. The number of races included in the analysis is 1,384 for males and 1,387 for females. For the creation of the Olympics-only database, only competitions with Major Game indication were considered. The total number of events per year is shown in Figure 1.

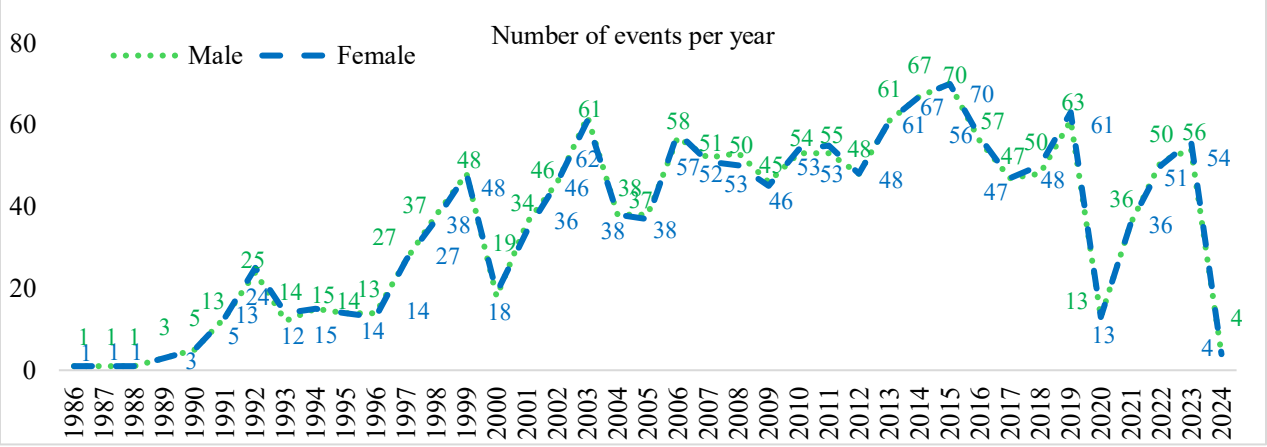


Figure 1. Number of events per year for male (A) and female (B) category.

Linear regression – Olympic Games

The average times of the top 3 finishers in each event at the Olympic Games are described in Table 1. The proportion of time for each discipline in the final result is illustrated in Figure 2. The prediction of the average of the top 3 times for Paris 2024, based on the data in Table 1 (excluding Paris 2024), is presented in Table 2.

Table 1. Average podium finishers time (s) at the Olympic Games.

		Swim	Bike	Run	Total
Sydney 2000	Male	1078	3527	1867	6515
	Female	1157	3915	2128	7249
Athens 2004	Male	1096	3628	1935	6677
	Female	1166	4122	2145	7494
Beijing 2008	Male	1098	3536	1850	6538
	Female	1192	3857	2048	7158
London 2012	Male	1022	3552	1760	6399
	Female	1161	3933	2022	7189
Rio de Janeiro 2016	Male	1044	3323	1885	6317
	Female	1151	3681	2078	7004
Tokyo 2020	Male	1090	3374	1783	6314
	Female	1110	3777	2026	6990
Paris 2024	Male	1249	3104	1790	6218
	Female	1356	3495	1971	6902
Mean ± SD	Male	1096.71 ± 72.97	3434.86 ± 180.48	1838.57 ± 63.26	6425.43 ± 159.10
	Female	1184.71 ± 79.36	3825.71 ± 200.39	2059.71 ± 61.69	7140.86 ± 198.98

Note. Mean ± SD = average time (s) and standard deviation of the discipline across all Olympic Games events.

Table 2. Podium finishers time estimation for Paris 2024. Linear Regression considering Olympic Games data.

		Swim	Bike	Run	Total
Slope	Male	-4.91	-47.54	-18.86	-63.54
	Female	-8.89	-55.34	-21.06	-78.11
Intercept	Male	1088.53	3656.40	1912.67	6682.40
	Female	1187.27	4074.53	2148.20	7454.07
R^2	Male	.22	.72	.37	.79
	Female	.19	.67	.72	.72
p-value	Male	<.001	<.001	<.001	<.001
	Female	<.001	<.001	<.001	<.001
Real Paris 2024 time (s)	Male	1249	3104	1790	6218
	Female	1356	3495	1971	6902
Estimated Paris 2024 time (s)	Male	1054	3324	1781	6237
	Female	1125	3687	2001	6907
Absolute difference (s)	Male	195	-220	9	-19
	Female	231	-192	-30	-5
Relative difference (%)	Male	18.49	-6.61	0.52	-0.30
	Female	20.53	-5.21	-1.49	-0.07

*Note. Intercept = Point where the regression line crosses the Y-axis when all predictor variables are zero; X = Model coefficient reflecting the impact of predictor variables on the estimated time; R^2 = Coefficient of determination; p-value = p-value (polynomial regression); Significance level = .05; Absolute difference (s) = Real time - Estimated time; Relative difference (%) = {[(Real time - Estimated time) / Estimated time] * 100}. Positive values indicate the predicted time is less than the real time, while negative values indicate the predicted time exceeds the real time.*

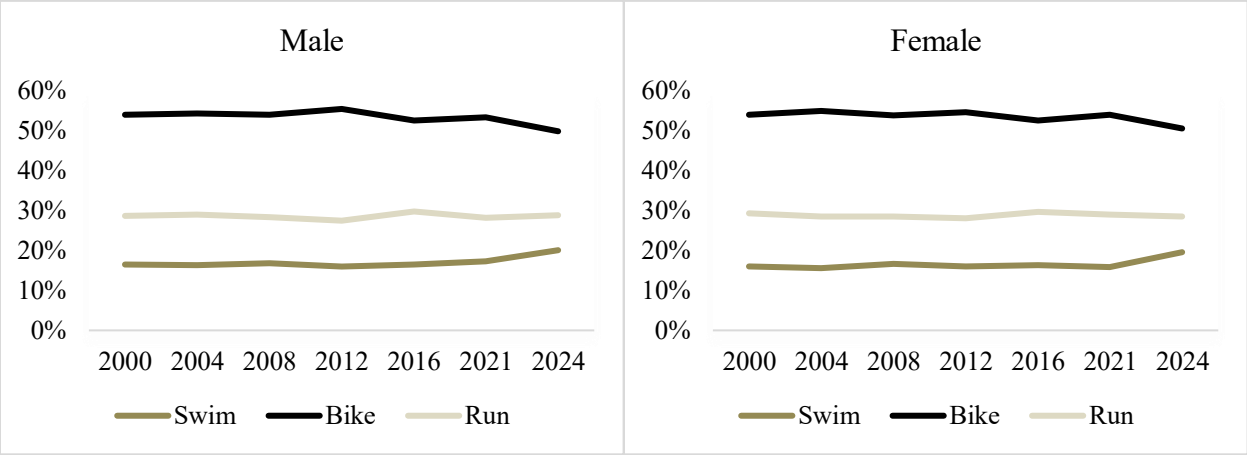


Figure 2. Discipline proportion (%) evolution for podium finishers at the Olympic Games.

When analysing previous Olympic editions, a trend of decreasing race times across the years can be identified (Table 2). For the predicted times, only the individual discipline times excluding transitions were considered, whereas the total time includes transitions (Table 2). Linear regression analysis indicates substantial differences in swim times (differences: 18.49% for males; 20.53% for females), moderate increased estimations in bike times (differences: -6.61% for males; -5.21% for females), and minimal differences in run times (differences: 0.52% for males; -1.49% for females). The prediction for total race time showed remarkable accuracy, with a difference of -0.30% for males and -0.07% for females (Table 2).

Linear regression – All data

Table 3. Podium finishers time estimation for Paris 2024. Linear Regression considering all data.

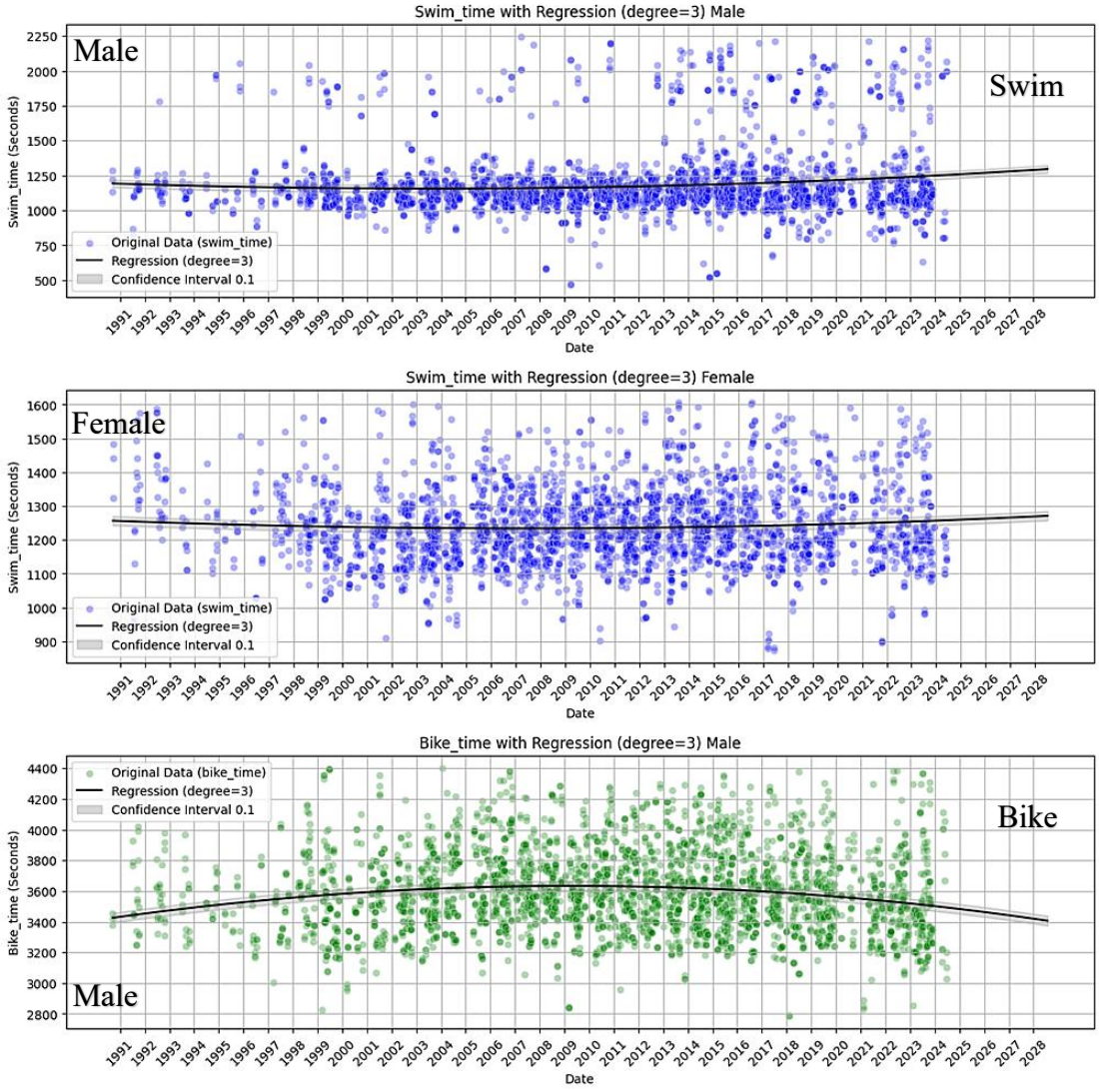
		Swim	Bike	Run	Total
Slope	Male	0.008	-0.003	-0.006	0.007
	Female	0.001	-0.004	0.001	0.006
Intercept	Male	-4745.554	5802.640	5989.906	1492.228
	Female	404.609	6773.265	1545.449	2998.050
R^2	Male	.009	.001	.003	.003
	Female	.001	.001	.0002	.001
p-value	Male	<.001	.074	.002	.003
	Female	.141	.079	.427	.052
Real Paris 2024 time (s)	Male	1249	3104	1790	6218
	Female	1356	3495	1971	6902
Estimated Paris 2024 time (s)	Male	1221	3582	1891	6791
	Female	1244	4001	2275	7627
Absolute difference (s)	Male	28	-478	-101	-573
	Female	112	-506	-304	-725
Relative difference (%)	Male	2.29	-13.34	-5.34	-8.44
	Female	9.00	-12.65	-13.36	-9.51

*Note. Slope = Change in the dependent variable (time) for each one-unit change in the independent variable (year); Intercept = Point where the regression line crosses the Y-axis when all predictor variables are zero; R^2 = Coefficient of determination; p-value = p-value (polynomial regression); Significance level = .05; Absolute difference (s) = Real time - Estimated time; Relative difference (%) = {[(Real time - Estimated time) / Estimated time] * 100}. Positive values indicate the predicted time is less than the real time, while negative values indicate the predicted time exceeds the real time.*

In the analysis of the time projections leading up to the Paris 2024 Olympic distance race, a total of 67 days passed between the most recent Olympic distance event and the Olympic Games race. Linear regression analysis indicates significant minimal decreased estimations in swim times for males (2.29%; *p*-value < .001), substantial increased estimations in run times for males (-5.34%; *p*-value < .001) and moderate increased estimations in total times for males (-8.44%; *p*-value < .001) in comparison with the real time (Table 3).

Third-degree polynomial regression

A third-degree polynomial regression was conducted to analyse the data trends. Figure 3 presents a scatter plot showing all considered data points alongside the resulting polynomial relationship. The graphs extend up to the year 2028, allowing for a clear visualization of the regression trend and its implications for future performance predictions. Once again, when analysing all the data, the same trend is confirmed as observed with the Olympic data alone: the times are projected to decrease in the future, particularly in the male category. The illustration shows a consistent pattern of performance improvement over time across various competitive events (Figure 3).



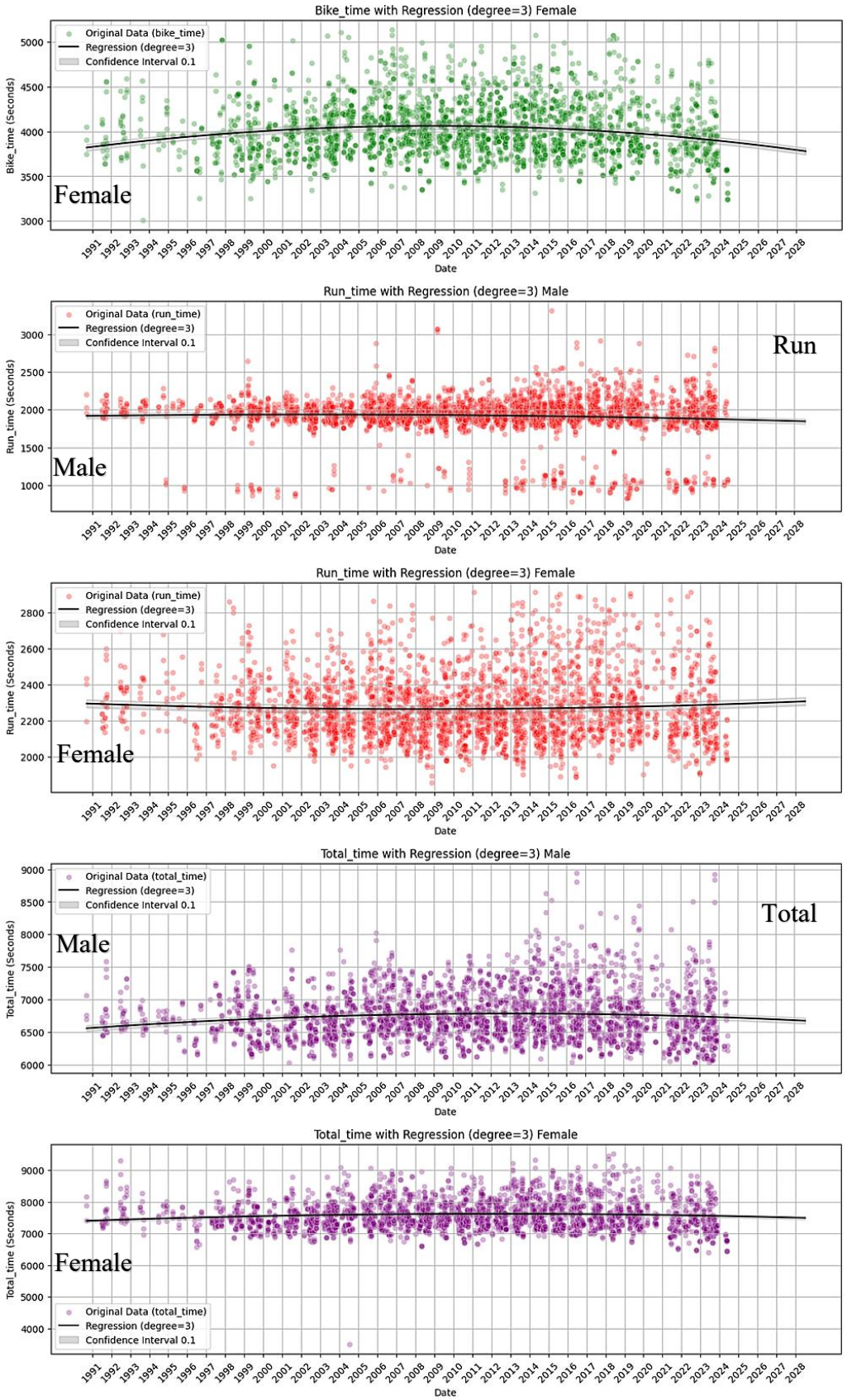


Figure 3. Podium finishers third-degree Polynomial Regression considering all data.

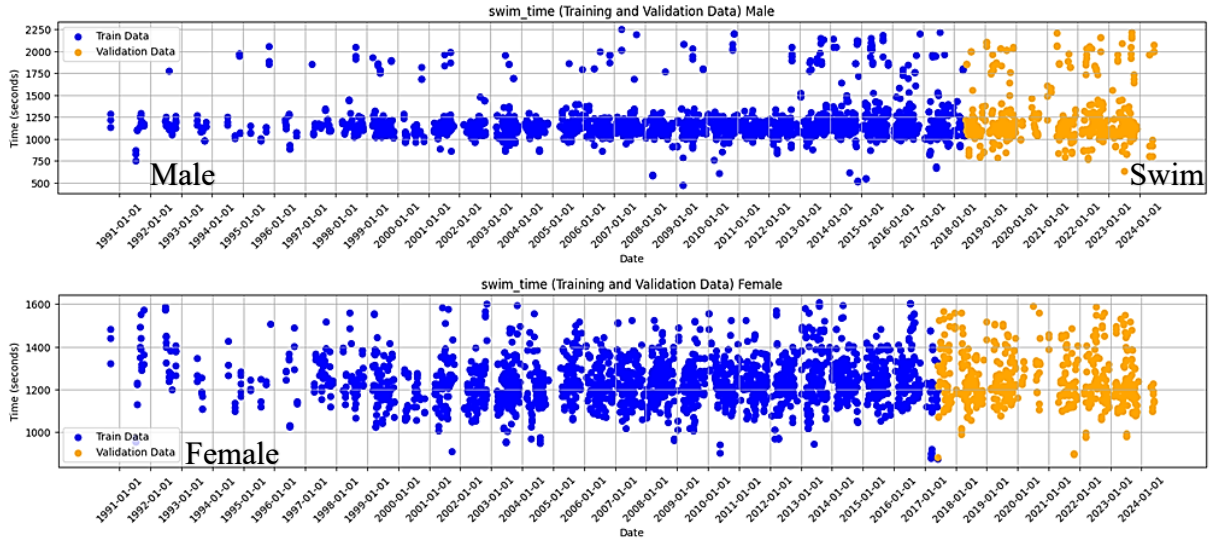
Table 4. Podium finishers time estimation for Paris 2024. Third-degree polynomial regression.

		Swim	Bike	Run	Total
Coefficients	Male	0, -2.27×10^{-12}, -1.66×10^{-6}, 1.52×10^{-12}	0, 6.21×10^{-12}, 4.55×10^{-6}, -4.14×10^{-12}	0, 1.33×10^{-12}, 9.79×10^{-7}, -8.93×10^{-13}	0, 4.70×10^{-12}, 3.45×10^{-6}, -3.13×10^{-12}
	Female	0, -8.46×10^{-13}, -6.20×10^{-7}, 5.64×10^{-13}	0, 7.53×10^{-12}, 5.52×10^{-6}, -5.02×10^{-12}	0, -1.06×10^{-12}, -7.75×10^{-7}, 7.52×10^{-13}	0, 4.98×10^{-12}, 3.65×10^{-6}, -3.31×10^{-12}
Intercept	Male	297805.64	-812835.31	-172398.29	-614134.78
	Female	112194.93	-985682.36	141185.98	64869.81
R^2	Male	.013	.024	.004	.009
	Female	.003	.025	.002	.006
p-value	Male	.017	<.001	<.001	<.001
	Female	.001	<.001	<.001	<.001
Real Paris 2024 time (s)	Male	1249	3104	1790	6218
	Female	1356	3495	1971	6902
Estimated Paris 2024 time (s)	Male	1253	3494	1872	6724
	Female	1257	3885	2291	7550
Absolute difference (s)	Male	-4	-390	-82	-506
	Female	99	-390	-320	-648
Relative difference (%)	Male	-0.32	-11.16	-4.38	-7.53
	Female	7.88	-10.04	-13.97	-8.58

*Note. Coefficients: Values associated with each term of the polynomial fitted to the data; Intercept = Point where the regression line crosses the Y-axis when all predictor variables are zero; R^2 = Coefficient of determination; p-value = p-value (polynomial regression); Significance level = .05; Absolute difference (s) = Real time - Estimated time; Relative difference (%) = {[(Real time - Estimated time) / Estimated time] * 100}. Positive values indicate the predicted time is less than the real time, while negative values indicate the predicted time exceeds the real time.*

Third-degree polynomial regression reveals that the estimated Olympic race times demonstrate slightly lower percentage differences when compared with the results from Table 3. Third-degree polynomial regression analysis indicates significant substantial and moderate decreased estimations in swim times (differences: 14.28% for males; 8.22% for females), moderate increased estimations in bike times (differences: -7.48% for males; -9.97% for females), moderate and substantial increased estimations in run times (differences: -6.12% for males; -13.63% for females) and minimal and moderate increased estimations in total times (differences: -3.53% for males; -8.34% for females) (Table 3).

PyTorch and TensorFlow analysis



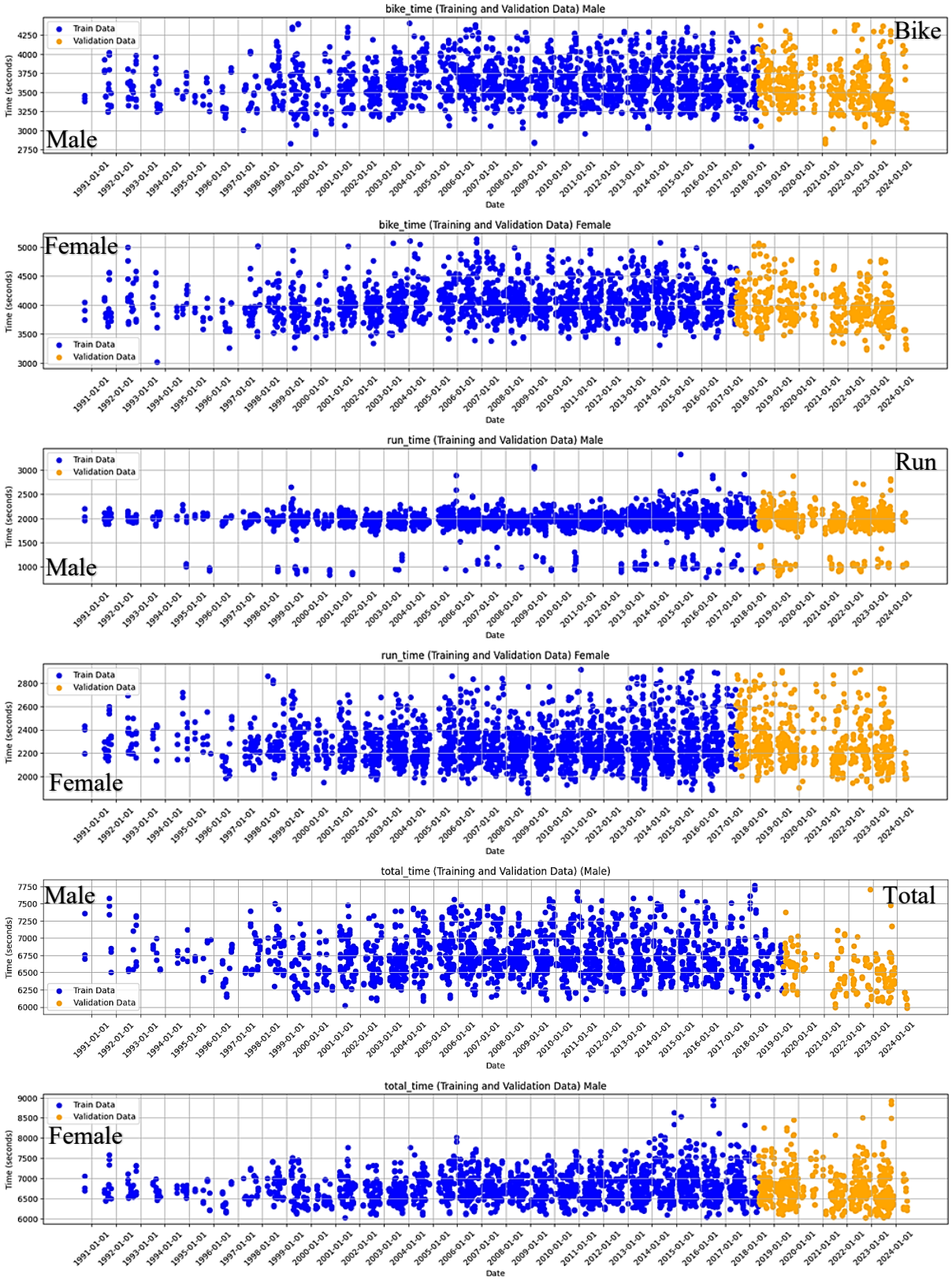


Figure 4. Podium finishers estimation. Training and validation data. Machine learning.

Table 4. Podium finishers time estimation for Paris 2024. Machine learning.

	Real Paris 2024 time (s)	Estimated Paris 2024 time (s)	Absolute difference (s)	Relative difference (%)
Males				
PyTorch				
Swim	1249	1204	45	3.74
Bike	3014	3409	-395	-11.59
Run	1790	1795	-5	-0.28
Total	6218	6499	-281	-4.32
TensorFlow				
Swim	1249	1333	-84	-6.30
Bike	3014	3457	-443	-12.81
Run	1790	1651	-139	-8.42
Total	6218	6442	-224	-3.48
Females				
PyTorch				
Swim	1356	1206	150	12.44
Bike	3495	3829	-334	-8.72
Run	1971	2195	-224	-10.21
Total	6902	7283	-381	-5.23
TensorFlow				
Swim	1356	1197	159	13.28
Bike	3495	3799	-304	-8.00
Run	1971	2174	-203	-9.34
Total	6902	7171	-269	-3.75

*Note. Absolute difference (s) = Real time - Estimated time; Relative difference (%) = {[(Real time - Estimated time) / Estimated time] * 100}. Positive values indicate the predicted time is less than the real time, while negative values indicate the predicted time exceeds the real time.*

Machine learning techniques were employed to analyse data trends. Figure 4 presents a scatter plot illustrating data points, with training data depicted in blue and predicted times represented in yellow. Both techniques (PyTorch and TensorFlow) showed comparable results in the male category, with bike times of -11.59% for PyTorch and -12.81% for TensorFlow, and total times of -4.32% for PyTorch and -3.48% for TensorFlow (Table 4). The average relative difference in the male category is -3.11% for PyTorch and -3.54% for TensorFlow, indicating a difference between them of 0.43% (Table 5). In the female category, all disciplines showed similar relative differences values, as well as the average relative difference (Table 4).

Overall summary

Run time was the best-predicted discipline for males (average difference: -0.21% ± 5.45%), and total time was the best-predicted variable for females (average differences: -5.43% ± 3.81%) (Table 5). Bike time exhibited the worst predictions for males (average difference: -11.10% ± 2.66%) whereas swim time produced the worst predictions for females (average difference: -12.63% ± 4.97%) (Table 5).

In the male category, the third-degree polynomial regression, linear regression considering Olympic Games races data, PyTorch and linear regression considering Olympic Games races data again were the most accurate predictions for swim (-0.32%), bike (-6.61%), run (-0.28%), and total times (-0.30%), respectively. Notably, TensorFlow was never the best technique for predicting any specific discipline.

In the female category, the third-degree polynomial regression was the best technique to predict the swim times (7.88%) whereas linear regression considering Olympic Games races data, was the most accurate for

bike (-5.21%), run (-1.49%), and total times (-0.07%). As in the male category, TensorFlow never emerged as the best technique for predicting any particular discipline (Table 5).

Table 5. Summary of relative differences (%) among the different prediction models.

	Conventional			Machine Learning		
	LR-OG (%)	LR-AD (%)	PR-3D (%)	PyTorch (%)	TensorFlow (%)	*Mean (%) ± SD
Male						
Swim	18.49	2.29	-0.32	3.74	-6.30	3.58 ± 9.18
Bike	-6.61	-13.34	-11.16	-11.59	-12.81	-11.10 ± 2.66
Run	0.52	-5.34	-4.38	-0.28	8.42	-0.21 ± 5.45
Total	-0.30	-8.44	-7.53	-4.32	-3.48	-4.81 ± 3.27
˚Mean ± SD	3.03 ± 10.79	-6.21 ± 6.56	-5.85 ± 4.61	-3.11 ± 6.54	-3.54 ± 8.88	
Female						
Swim	20.53	9.00	7.88	12.44	13.28	12.63 ± 4.97
Bike	-5.21	-12.65	-10.04	-8.72	-8.00	-8.92 ± 2.73
Run	-1.49	-13.36	-13.97	-10.21	-9.34	-9.67 ± 4.99
Total	-0.07	-9.51	-8.58	-5.23	-3.75	-5.43 ± 3.81
˚Mean ± SD	3.44 ± 11.60	-6.63 ± 10.55	-6.18 ± 9.64	-2.93 ± 10.46	-1.95 ± 10.43	

*Note. LR-OG = Linear Regression using only Olympic Games data; LR-AD = Linear Regression using all available data; PR-3D = Third-degree polynomial Regression using all data; PyTorch = Machine Learning Neural Network implemented with PyTorch; TensorFlow = Machine Learning Neural Network implemented with TensorFlow; Mean (SD) = Average relative change (%) of the model (row) or discipline (column), with standard deviation. Relative change was calculated as {[(real time - predicted time) / predicted time] * 100}. Positive values indicate the predicted time is less than the real time, while negative values indicate the predicted time exceeds the real time.*

**: Mean average and standard deviation of the different disciplines.*

˚: Mean average and standard deviation of the different estimations of the different statistical techniques.

For males, the linear regression based on Olympic Games races data was the most accurate technique overall, as it showed the lowest average relative difference (-3.03% ± 10.79%), and providing the most precise predictions in two disciplines (bike and total times), although all methods showed acceptable average accuracy. In the female category the linear regression using Olympic Games race data exhibited the most accurate prediction across most disciplines (bike, run and total times), but TensorFlow achieved the best precision (average relative difference: -1.95% ± 10.43%). All variables were better predicted by the same statistical techniques in both sexes except for run time, that was better predicted by PyTorch for males (-0.28%) and by linear regression based on Olympic Games races data for females (-1.49%).

DISCUSSION

The aim of the present study was to predict triathlon performance at the Paris 2024 Olympics using conventional statistics and a Machine Learning-based approach. The secondary aim of this study was to compare machine learning methods and conventional statistical techniques to evaluate their precision and accuracy in predicting triathlon performance outcomes. It was hypothesized that both predictive models would grant sufficiently accurate results and that machine learning methods could be more accurate than conventional statistical techniques. General statistical methods and machine learning techniques have proven to be reliable approaches for predicting performance in Olympic triathlon, especially when estimating the running performance for males and the total time for women. For males, the linear regression based on Olympic Games races data was the best predictable technique (average difference: -3.03% ± 10.79%), followed closely by machine learning-based technique (average differences: -3.11% ± 6.54%; 3.54% ± 8.88% for PyTorch and TensorFlow respectively). For females, the TensorFlow machine learning-based

technique was the most accurate on average (average difference: -1.95% ± 10.43%), although the linear regression based on Olympic Games races data was the technique that predicted best most variables.

Linear regression analysis considering data from previous Olympic Games events indicates substantial differences between predicted and actual performance times in the swimming segment. Indeed, in swimming, the differences were 18.49% for males and 20,53% for females, indicating an overestimation of the time to complete the section. Conversely, the cycling section shows a moderate prediction with differences of -6.61% for males and -5,21% in females. Similarly, the running prediction shows a good agreement between the prediction and the actual performance times (differences: 0.52% for males; -1.49% for females). The prediction for total race time showed remarkable accuracy, with a difference of -0.30% for males and -0.07% for females (Table 2). Considering the accuracy results reported by other authors (Lerebourg et al., 2022; Lim & Song, 2024), this technique demonstrates comparable accuracy levels for cycling and running times, higher accuracy for total race time, and lower accuracy for swimming time predictions.

The considerable difference observed in the swimming leg can likely be attributed to the unique conditions experienced in Paris, particularly the swimming course in the river, which differed from those of previous editions. These conditions may have influenced the athletes' performance, either through environmental factors or psychological effects. Notably, in the cycling segment, a small difference occurred despite the fact that the circuit featured negligible elevation gain, unlike previous editions. In the running leg, the reduced difference may be attributed to the assumption that the running course was the most consistent across Olympic editions, typically being flat, and this database only considered data from previous Olympic Games Finally, the minimal difference in total race time highlights the potential of this predictive approach to reliably forecast outcomes in future editions.

The results are different when extending the analysis to include all data in the database (i.e., ETU, WC, WTS). Linear regression analysis indicates minimal and moderate overestimation of swim times (differences: 2.29% for males; 9.00% for females), substantial underestimations in bike times (differences: -13.34% for males; -12.65% for females), moderate and substantial underestimations in run times (differences: -8.44% for males; -13.36% for females) and moderate underestimations in total times (differences: -8.44% for males; -9.51% for females) (Table 3). The discrepancies between predictions may be attributed to 1) differences in the database of each sex, as historically there have been fewer female participants in Olympic triathlon events compared to male participants, and 2) performance differences between sexes, a factor that has been extensively explored in the literature. (García-González & González-Jurado, 2025b; Lepers et al., 2014; Lepers & Stapley, 2010; Piacentini et al., 2019).

Third-degree polynomial regression analysis indicates an accurate prediction in men's swim times and a moderate overestimation in the female category (differences: -0.32% for males; 7.88% for females), substantial underestimations in cycling times (differences: -11.16% for males; -10.04 % for females), minimal and substantial underestimations in run times (differences: -4.38% for males; -13.97% for females) and moderate underestimations in total times (differences: -5.85% for males; -6.18% for females) (Table 4).

Liew et al. (2022) advocated for caution in the optimism of applying machine learning in prognostic modelling, and its benefit is likely dependent on factors like sample size, variable type, disease investigated, to name a few. This statement aligns with our findings. While TensorFlow may be observed as the best overall predictor of performance in the female category, it should not be applied indiscriminately. Specifically, for the male category, the linear regression analysis considering data from previous Olympic Games demonstrated the highest accuracy among all techniques, despite all methods showing acceptable performance (Table 5).

Likewise, in the female category, the linear regression based on Olympic Games races data was the technique that predicted best most variables. These findings suggest that no single technique can be considered clearly superior. Instead, the optimal predictive model depends on the sex category, the structure and availability of the data, and the specific outcome being predicted. Consequently, model selection should be driven by empirical performance under clearly defined conditions, instead of by a general preference for more complex machine learning approaches.

Run was the best-predicted discipline with all methods used for the male category (-0.21% ± 5.45%) whereas total time was the best-predicted discipline for the female category (-5.43% ± 3.81%). Bike time exhibited the worst predictions for males (-11.10% ± 2.66%) and swim time was the worst predicted variable for females (12.63% ± 4.97%). The estimated swim times were lower than the actual event times, likely due to the unique environmental features. TensorFlow was never the best technique for predicting any specific discipline in any sexes, although it showed the best average difference in the female category. Although the linear regression analysis using data from previous Olympic Games was the most accurate statistical technique, it should be noted that the sample size used for this method was considerably smaller than that of the other techniques. Therefore, when comparing statistical methods based on the same dataset, a slight improvement in prediction accuracy can be observed with artificial intelligence-based techniques.

Both techniques utilized (PyTorch and TensorFlow) showed nearly identical average results for both sexes (Table 5). The total times presented a difference of -4.32% for males and -5.23% for females in PyTorch, and -3.48% for males and -3.75% for females in TensorFlow (Table 5). To contextualize these results and compare them with other studies utilizing Machine Learning for event prediction, Nagovitsyn et al. (2023) reported an 11% error probability in predicting a wrestler's competitive performance. Moreover, Lim & Song (2024) presented a model that explained approximately 93% of the variability in clean & jerk performance. Additionally, Lerebourg et al. (2022) achieved a prediction accuracy above 94% for marathon performance, with the KNN model performing better than ANN, reaching an accuracy above 98%. Lastly and focusing on triathlon, other authors reported a model accuracy of 81.25% in predicting triathlon mixed relay successful race positions from their validation sample (Martínez-Sobrino et al., 2023). The largest total time difference between the estimated time predicted by our models and the real time achieved is -8.58%, which demonstrates similar accuracy compared to the aforementioned studies.

The present study has some limitations. This study adopts a quantitative approach, which implies a lack of consideration for physical or other nature-related variables, such as training, previous experience, circuit layout or climatic conditions. Finally, analysing together competitions over a long period, including a small number of competitions where drafting was allowed and competitions where drafting was forbidden may have a small influence on the results, although preliminary analysis show that the results would be similar to those exposed in this study.

This study has practical implications for training in Olympic distance triathlon. This insight provides an indication of where the sport of triathlon is heading and may help refine future race strategies. Furthermore, these findings can be applied when setting performance targets for major events, allowing for more precise goal-setting and preparation. The model's results showed a considerable degree of reliability. For context, predictive accuracy in meteorology for a 7-day forecast ranges between 76% and 87%, while predictions extending beyond this period decrease to an accuracy range of 31 to 40% (Chen et al., 2023). Several other factors, including genetic predisposition, environmental conditions, access to technological advancements, and evolving training techniques, play pivotal roles in influencing an athlete's performance.

CONCLUSION

In conclusion, general statistical methods and machine learning techniques have proven to be reliable approaches for predicting performance in Olympic triathlon. PyTorch and TensorFlow showed better predictions than conventional statistics. For men, the linear regression based on Olympic Games races data was the most accurate technique overall. For women, TensorFlow achieved the best precision but the linear regression using Olympic Games race data exhibited the most accurate prediction across most disciplines. The estimated times were almost always higher than the actual times in Paris 2024, except in the swimming discipline, probably due to specific conditions.

AUTHOR CONTRIBUTIONS

All authors meet the criteria for authorship in accordance with established ethical guidelines. Luca A. Bianchini was responsible for data handling and analyses, and contributed to investigation, writing, and editing. Pablo García González focused on conceptualization, investigation, drafting, and visualization. Andrea Fuk and Simone Villanova contributed to review and editing. José Antonio González Jurado contributed to validation, resources, supervision, and editing. Maria Francesca Piacentini contributed to validation, editing, supervision, and project administration. All authors have critically reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

AI USE DISCLOSURE

In accordance with current publishing ethics and transparency recommendations, artificial intelligence (AI) tools were used solely to assist with translation and language editing, with the aim of improving clarity and readability. No AI tools were used in the generation of scientific content, including the study design, data collection, analysis, interpretation of results, or the formulation of conclusions. The authors retain full responsibility for the content of the manuscript and confirm its originality, integrity, and accuracy.

REFERENCES

Bullock, G. S., Mylott, J., Hughes, T., Nicholson, K. F., Riley, R. D., & Collins, G. S. (2022). Just How Confident Can We Be in Predicting Sports Injuries? A Systematic Review of the Methodological Conduct and Performance of Existing Musculoskeletal Injury Prediction Models in Sport. Sports Medicine, 52(10), 2469-2482. https://doi.org/10.1007/s40279-022-01698-9

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. Science and Engineering Ethics, 24(2), 505-528. https://doi.org/10.1007/s11948-017-9901-7

Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., & Yang, Z. (2023). Machine Learning Methods in Weather and Climate Applications: A Survey. Applied Sciences, 13(21), 12019. https://doi.org/10.3390/app132112019

Claudino, J. G., Capanema, D. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., & Nassis, G. P. (2019). Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. Sports Medicine - Open, 5(1), 1-12. https://doi.org/10.1186/s40798-019-0202-3

Cuba-Dorado, A., Vleck, V., Álvarez-Yates, T., & Garcia-Garcia, O. (2021). Gender effect on the relationship between talent identification tests and later world triathlon series performance. Sports, 9(12), 164. https://doi.org/10.3390/sports9120164

Dindorf, C., Bartaguiz, E., Gassmann, F., & Fröhlich, M. (2023). Conceptual Structure and Current Trends in Artificial Intelligence, Machine Learning, and Deep Learning Research in Sports: A Bibliometric Review. International Journal of Environmental Research and Public Health, 20(1), 173. https://doi.org/10.3390/ijerph20010173

García-González, P., Bianchini, L. A., Fuk, A., Villanova, S., González-Jurado, J. A., & Piacentini, M. F. (2026). Applying Artificial Intelligence to Determine the Required Positions in Each Discipline for Overall Olympic Triathlon Success. Applied Sciences, 16(6), 2871. https://doi.org/10.3390/app16062871

García-González, P., & González-Jurado, J. A. (2024a). Has COVID-19 influenced the performance of top-class athletes in the ITU World Duathlon and World Aquathlon Championship? Revista Andaluza de Medicina Del Deporte, 17(3), 151-158. https://doi.org/10.33155/ramd.v17i3-4.1177

García-González, P., & González-Jurado, J. A. (2024b). Has Covid-19 pandemic influenced the performance of top-class triathletes in the Sprint and Olympic distance of the ITU World Triathlon Championship Series? Retos, 57, 137-146. https://doi.org/10.47197/retos.v57.105394

García-González, P., & González-Jurado, J. A. (2025a). Analysis and Comparison of Female Triathlon Top-Class Performance at Three Main Competitions over 23 Years Considering Race Position. Annals of Applied Sport Science, 13(2), e1490. https://doi.org/10.61186/aassjournal.1490

García-González, P., & González-Jurado, J. A. (2025b). Analysis and comparison of male triathlon performance at the Olympic Games, International Triathlon Union World Championship, and European Triathlon Union European Championship over a period of twenty-three years, considering both medalists and overall triathletes participants. Gazzetta Medica Italiana - Archivio per Le Scienze Mediche, 184(10), 835-843. https://doi.org/10.23736/S0393-3660.24.05732-2

Knechtle, B., Thuany, M., Valero, D., Villiger, E., Nikolaidis, P. T., Andrade, M. S., Cuk, I., Rosemann, T., & Weiss, K. (2025). The association of origin and environmental conditions with performance in professional IRONMAN triathletes. Scientific Reports, 15(1), 2700. https://doi.org/10.1038/s41598-025-86033-8

Lepers, R., Knechtle, B., & Stapley, P. (2014). Trends in Triathlon Performance: Effects of Sex and Age. Sports Medicine, 43(9), 851-863. https://doi.org/10.1007/s40279-013-0067-4

Lepers, R., & Stapley, P. J. (2010). Differences in gender and performance in off-road triathlon. Journal of Sports Sciences, 28(14), 1555-1562. https://doi.org/10.1080/02640414.2010.517545

Lerebourg, L., Saboul, D., Clémençon, M., & Coquart, J. B. (2022). Prediction of Marathon Performance using Artificial Intelligence. International Journal of Sports Medicine, 44(5), 352-360. https://doi.org/10.1055/a-1993-2371

Liew, B. X. W., Kovacs, F. M., Rügamer, D., & Royuela, A. (2022). Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. European Spine Journal, 31(8), 2082-2091. https://doi.org/10.1007/s00586-022-07188-w

Lim, B., & Song, W. (2024). Exploring CrossFit Performance Prediction and Analysis via Extensive Data and Machine Learning. The Journal of Sports Medicine and Physical Fitness, 64(7), 640-649. https://doi.org/10.23736/S0022-4707.24.15786-6

Malcata, R. M., Hopkins, W. G., & Pearson, S. N. (2014). Tracking career performance of successful triathletes. Medicine and Science in Sports and Exercise, 46(6), 1227-1234. https://doi.org/10.1249/MSS.0000000000000221

Martínez-Gramage, J., Albiach, J. P., Moltó, I. N., Amer-Cuenca, J. J., Moreno, V. H., & Segura-Ortí, E. (2020). A random forest machine learning framework to reduce running injuries in young triathletes. Sensors, 20(21), 6388. https://doi.org/10.3390/s20216388

Martínez-Sobrino, J., Veiga, S., Del Cerro, J. S., & González-Ravé, J. M. (2023). What is the Most Important Leg and Discipline in Triathlon Mixed Team Relays? Journal of Human Kinetics, 89, 269-278. https://doi.org/10.5114/jhk/167088

Nagovitsyn, R. S., Valeeva, R. A., & Latypova, L. A. (2023). Artificial Intelligence Program for Predicting Wrestlers' Sports Performances. Sports, 11(10), 196. https://doi.org/10.3390/sports11100196

Ofoghi, B., Zeleznikow, J., Macmahon, C., Rehula, J., & Dwyer, D. B. (2016). Performance analysis and prediction in triathlon. Journal of Sports Sciences, 34(7), 607-612. https://doi.org/10.1080/02640414.2015.1065341

Olaya-Cuartero, J., Fernández-Sáez, J., Østerlie, O., & Ferriz-Valero, A. (2022). Concordance Analysis between the Segments and the Overall Performance in Olympic Triathlon in Elite Triathletes. Biology, 11(6), 902. https://doi.org/10.3390/biology11060902

O'Toole, M. L., & Douglas, P. S. (1995). Applied Physiology of Triathlon. Sports Medicine, 19(4), 251-267. https://doi.org/10.2165/00007256-199519040-00003

Piacentini, M. F., Vleck, V., & Lepers, R. (2019). Effect of age on the sex difference in Ironman triathlon performance. Movement and Sports Sciences - Science et Motricite, (104), 21-27. https://doi.org/10.1051/sm/2019030

Reis, F. J. J., Alaiti, R. K., Vallio, C. S., & Hespanhol, L. (2024). Artificial intelligence and Machine Learning approaches in sports: Concepts, applications, challenges, and future perspectives. Brazilian Journal of Physical Therapy, 28(3), 101083. https://doi.org/10.1016/j.bjpt.2024.101083

Rossi, A., Pappalardo, L., & Cintia, P. (2021). A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer. Sports, 10(1), 5. https://doi.org/10.3390/sports10010005

Rothschild, J. A., Stewart, T., Kilding, A. E., & Plews, D. J. (2024). Predicting daily recovery during long-term endurance training using machine learning analysis. European Journal of Applied Physiology, 124(11), 3279-3290. https://doi.org/10.1007/s00421-024-05530-2

Ruiz-Tendero, G., & Salinero Martín, J. J. (2012). Psycho-Social Factors Determining Success In High-Performance Triathlon: Compared Perception In The Coach-Athlete Pair. Perceptual & Motor Skills: Physical Development & Measurement, 115(3), 865-880. https://doi.org/10.2466/08.25.PMS.115.6.865-880

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S. C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLoS Computational Biology, 15(7), e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Sousa, C. V., Aguiar, S., Olher, R. R., Cunha, R., Nikolaidis, P. T., Villiger, E., Rosemann, T., & Knechtle, B. (2021). What Is the Best Discipline to Predict Overall Triathlon Performance? An Analysis of Sprint, Olympic, Ironman® 70.3, and Ironman® 140.6. Frontiers in Physiology, 12. https://doi.org/10.3389/fphys.2021.654552

Sperlich, B., Düking, P., Leppich, R., & Holmberg, H. C. (2023). Strengths, weaknesses, opportunities, and threats associated with the application of artificial intelligence in connection with sport research, coaching, and optimization of athletic performance: a brief SWOT analysis. Frontiers in Sports and Active Living, 5, 1258562. https://doi.org/10.3389/fspor.2023.1258562

Tam, C. K., & Yao, Z. F. (2024). Advancing 100m sprint performance prediction: A machine learning approach to velocity curve modeling and performance correlation. PLoS ONE, 19(5), e0303366. https://doi.org/10.1371/journal.pone.0303366

Thuany, M., Valero, D., Villiger, E., Fernandes, M. S. S., Forte, P., Weiss, K., Nikolaidis, P. T., Cuk, I., & Knechtle, B. (2024). A study of the fastest courses for professional triathletes competing in IRONMAN® triathlons. Human Movement, 25(2), 148-160. https://doi.org/10.5114/hm/189332

Thuany, M., Valero, D., Villiger, E., Forte, P., Weiss, K., Nikolaidis, P. T., Andrade, M. S., Cuk, I., Sousa, C. V., & Knechtle, B. (2023). A Machine Learning Approach to Finding the Fastest Race Course for Professional Athletes Competing in Ironman® 70.3 Races between 2004 and 2020. International Journal of Environmental Research and Public Health, 20(4), 3619. https://doi.org/10.3390/ijerph20043619

Van Schuylenbergh, R., Eynde, B. V., & Hespel, P. (2004). Prediction of sprint triathlon performance from laboratory tests. European Journal of Applied Physiology, 91(1), 94-99. https://doi.org/10.1007/s00421-003-0911-6

Weiss, K., Valero, D., Andrade, M. S., Villiger, E., Thuany, M., & Knechtle, B. (2024). Cycling is the most important predictive split discipline in professional Ironman® 70.3 triathletes. Frontiers in Sports and Active Living, 6, 1214929. https://doi.org/10.3389/fspor.2024.1214929

