SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

A Systematic Literature Review of Machine Learning for Endurance Running Performance Prediction

Efraim William Solang[1)*], Linawati[2)], Ida Bagus Gede Manuaba[3)], I Nyoman Setiawan[4)]
[1,2,3,4)]Universitas Udayana, Indonesia
[1)]efraimwilliam@gmail.com, [2)]linawati@unud.ac.id, [3)] ibgmanuaba @unud.ac.id, [4)]setiawan@unud.ac.id

Submitted :Dec 22, 2025 | **Accepted** : Jan 4, 2026 | **Published** : Jan 05, 2026

Abstract: This study systematically reviews the application of machine learning methods for predicting running performance, with particular emphasis on short-middle distance events such as the 5 km. Although machine learning based performance prediction has been widely explored in endurance sports, a comprehensive review synthesizing models, predictors, and pipelines across running distances remains limited. The review followed the PRISMA 2020 framework. Articles published between 2020 and 2025 were retrieved from ScienceDirect, Google Scholar, and PubMed using predefined keyword combinations related to machine learning and running performance. Studies were included if they focused on running (excluding cycling, triathlon, or other sports), applied predictive modeling, and reported model evaluation metrics. A total of 26 studies met the inclusion criteria and were assessed using quality appraisal criteria inspired by TRIPOD and QUADAS-2. The analysis identified four main research themes: (1) application of machine learning models for running performance prediction, (2) physiological and anthropometric predictors, (3) non-physiological and contextual factors, and (4) personalized athlete training and monitoring. Ensemble learning models (Random Forest, XGBoost, LightGBM) consistently outperformed traditional linear regression by capturing non-linear interactions, while deep learning approaches (LSTM, GRU) demonstrated strong capability in modeling temporal training dynamics. A generalized machine learning pipeline for running performance prediction was also synthesized. This review contributes a structured framework that integrates modeling approaches, predictor categories, and evaluation strategies, and highlights research opportunities for explainable and personalized prediction systems, particularly for 5 km running performance.

Keywords: Machine learning; prediction of running time; running performance; sports analytics; systematic literature review

INTRODUCTION

This Running has seen a significant surge in popularity in recent years, transitioning from a mere fitness activity to a modern lifestyle. Indonesia. In 2021, Indonesia was among the top five countries with the highest diabetes rates in the world. This may be a reason why people are starting to adopt a healthy lifestyle, particularly exercise, especially running (Winandy Soesilo & Suprapti Hendharta, 2025). According to Garmin's official website, the trend of outdoor running has seen a significant increase of 65% (Garmin Indonesia, 2024). This demonstrates the growing public interest in running, which can improve cardiorespiratory fitness, muscle strength, flexibility, and emotional health. This increase is reflected not only in the increasing number of runners but also in the numerous running events and competitions held throughout the year. Typical running competitions are typically 5 km. Results are determined by time, making predicting running times crucial for runners.

Running performance prediction has become an important topic in the field of sports science and physiological data analysis because the prediction results will help coaches and athletes in planning optimal training strategies (Nikolaidis et al., 2021; Wiecha et al., 2022). Predictions can be made using machine learning technology. The resulting running data can be used as a machine learning resource to model the complex relationship between training variables and race times.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In the last five years, machine learning has become a superior alternative for modeling complex relationships in athletic performance. Studies have shown that the use of ensemble learning algorithms (XGBoost, LightGBM, Random Forest) has been shown to improve accuracy by 20% compared to conventional models (Rothschild et al., 2024), while neural network models are able to capture temporal patterns from sequential training data with up to 90% accuracy (Zhou, 2025). Furthermore, advances in wearable sensors and big data analytics have expanded the scope of analyzable variables, including heart rate, physiological variability, sleep patterns, and gait speed, which are relevant for personalized predictive models (Coquart, 2023; Nikolaidis & Knechtle, 2023)

Several existing reviews have addressed running performance modeling from physiological and training perspectives. (Alvero-Cruz et al., 2020) presented a narrative review of physiological and anthropometric factors in predicting long-distance running performance, while (Vijay et al., 2024) emphasized the role of lactate threshold-based training in improving endurance performance. While these studies provide an important foundation for understanding the determinants of running performance, their primary focus has been limited to conventional statistical approaches or training interventions, and they have not systematically explored the role of machine learning as a comprehensive, data-driven prediction framework.

To the best of our knowledge, there is still limited systematic literature review specifically synthesizing how machine learning algorithms, modeling pipelines (including preprocessing, feature engineering, and evaluation), and personalization approaches are used to predict long-distance running performance, with a particular emphasis on the 5 km event. Of the 26 studies that met the inclusion criteria in this review, only a small fraction explicitly focused on 5 km performance, while the majority addressed half-marathons, marathons, and ultramarathons. Therefore, this study aims to fill this gap by presenting a PRISMA-based systematic review that identifies key research themes, compares the machine learning approaches used, and formulates a conceptual framework for developing predictive models and personalizing running performance, specifically in the context of middle-distance running.

Therefore, a comprehensive literature review is needed to analyze the use of machine learning in the context of predicting running performance based on both physiological and non-physiological variables. This review will also compare the strengths and weaknesses of each model and identify opportunities for further research development.

METHOD

This study employed a Systematic Literature Review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020). The objective was systematically identified, evaluate, and synthesize empirical studies that applied machine learning or predictive modeling approaches to running performance prediction, with particular emphasis on the 5 km distance within the broader endurance running context. The research flow is illustrated in the diagram in Figure 1 below.

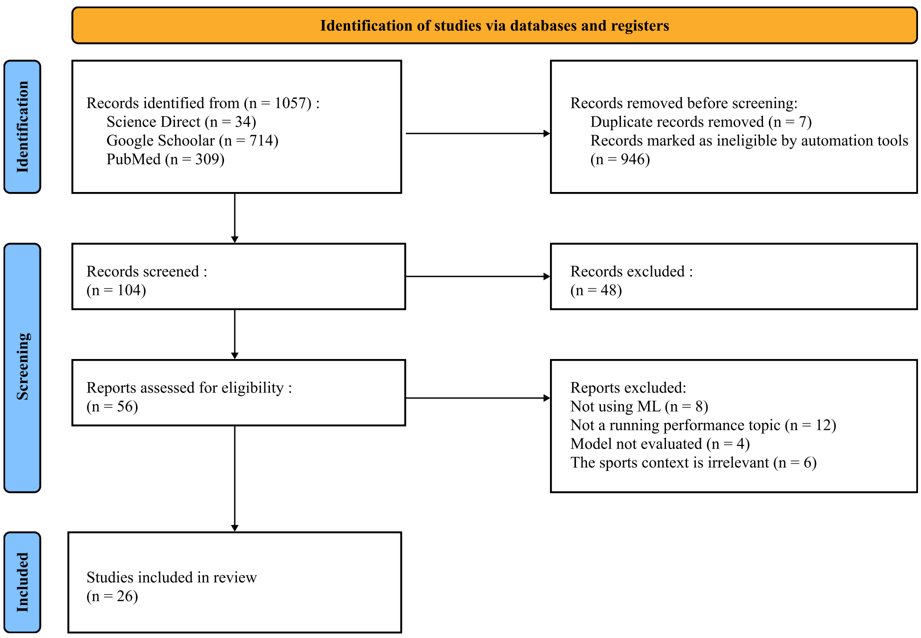

Fig. 1 Research Method Flow Systematic Literature Review (PRISMA)

The study selection followed a three-stage PRISMA process: identification, screening, and eligibility. The identification phase collected 1,057 scientific articles published between 2020 and 2025 through several major

*name of corresponding author

academic databases, namely ScienceDirect and Google Scholar. Keyword combinations used included "machine learning running performance" and "5 km running prediction". After an initial screening process, 7 duplicate articles and 946 articles were excluded using automation tools. This process resulted in 104 scientific articles deemed potentially relevant to the research topic [2].

On screening stage, 48 articles were excluded because they did not involve machine learning approaches or were not related to running activities, resulting in 56 retained articles. Further full-text eligibility assessment excluded 30 articles due to unmet criteria [3]. Consequently, 26 studies met all eligibility criteria and were included in the final qualitative synthesis.

The inclusion and exclusion criteria applied at each PRISMA stage are detailed below to ensure consistency between the narrative and the PRISMA flow diagram. The inclusion and exclusion criteria applied at each PRISMA stage are detailed below to ensure consistency between the narrative and the PRISMA flow diagram.

Inclusion Criteria:

- Studies applying machine learning or data-driven prediction models.
- Focus on individual running performance (e.g., 5 km, half-marathon, marathon, ultramarathon).
- Use of physiological, anthropometric, training, or contextual variables related to running.
- Reporting quantitative model evaluation metrics (e.g., R^2, RMSE, MAE, MAPE).
- Peer-reviewed articles published between 2020 and 2025.

Exclusion Criteria:

- Studies focusing on non-running sports (e.g., cycling, triathlon, or other sports).
- Articles not involving predictive modeling or machine learning approaches.
- Studies without quantitative performance evaluation results.
- Research addressing team-based performance rather than individual runners.

Research Theme

A thematic analysis was conducted on 26 articles that met the inclusion criteria. This process was conducted to identify research patterns and the direction of research development on the topic of machine learning in predicting 5 km running performance (Vijay et al., 2024). Figure 2 is the theme of the literature review that has been determined.

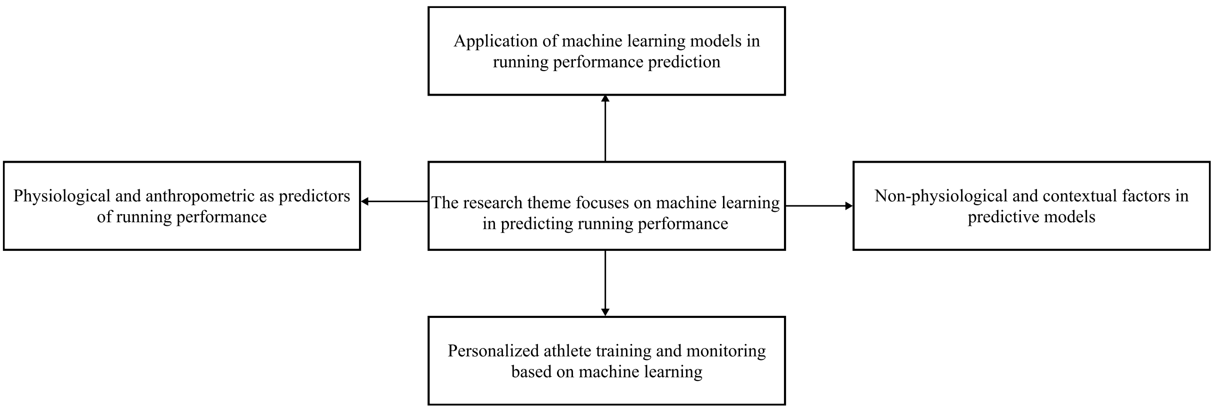

Fig. 2 Thematic Analysis

This study inductively identified four main themes through in-depth literature coding to ensure the accuracy of the results. The researchers acknowledge the potential limitations of subjectivity stemming from professional backgrounds and technical assumptions related to machine learning. However, to maintain objectivity, the entire analysis process focused on conceptual consistency, empirical relevance, and logical relationships between running performance variables. This approach ensures that the resulting thematic findings are representative of the current state of the literature and have strong academic accountability as a reference for future predictive model development.

RESULT

Following the PRISMA selection process, a total of 26 studies were included in the final analysis. To ensure methodological rigor and transparency, a structured quality appraisal was conducted for each study prior to synthesis.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Each study was evaluated based on a quality score (Q1-Q8), which aims to assess the methodological strength and quality of the research reports of studies that have passed the eligibility stage. This stage was conducted to ensure that only valid, transparent, and relevant studies were included. The quality scoring analysis was performed using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) methods [4], [5].

It is important to note that the TRIPOD and QUADAS-2 frameworks were not applied as formal diagnostic or reporting standards in this review. Instead, key principles from both guidelines were adapted and operationalized into a structured checklist (Q1–Q8) to assess the methodological transparency and rigor of machine learning based on running performance prediction studies. An explanation of each assessment criterion is shown in Table 1 below.

Table 1. TRIPOD Assessment Criteria

No	Criteria
1	The research objectives are aligned with the topic of performance prediction.
2	The clarity and completeness of the data used.
3	The variables used are relevant to running performance.
4	The machine learning algorithm is explained in detail.
5	The quality of the machine learning model evaluation.
6	The model results are explained in the context of running performance.
7	The research results can be generalized to other contexts.
8	The research results can be applied by runners and coaches.

Each criterion is divided into three parts (0 = poor, 1 = moderate, 2 = excellent). Studies with a score of 13–16 are considered excellent. Studies with a score of 9–12 are considered moderate, and scores below 9 indicate poor studies. Based on this assessment, 12 articles were classified as excellent, 14 as moderate, and 5 as poor. Table 2 provides a comparative overview of methodological rigor, predictive variables, and machine learning approaches across the reviewed literature using the QUADAS-2 method.

Tabel 2. QUADAS-2 and Core Characteristic

Author & Year	ML Methods	Evaluation Model	Result	Quality
Thuany et al., 2025	XGBoost	MAE & R^2	Gender was the strongest predictor, followed by nationalities, age, and race location. The fastest runners were predominantly from European countries.	Excellent
Weiss et al., 2024)	OLS & XGBoost	R2	Increased temperature and decreased humidity decreased running speed, with a greater effect on men. Sunshine duration was positively correlated but small.	Moderate
Zhou, 2025	Linear Regression, RNN, LSTM & GRU	RMSE, MAE, MAPE	The GRU model provided the most accurate and stable marathon time predictions compared to Linear Regression.	Moderate
Nikolaidis et al., 2021	Linear Regression	R^2	VO_2max, weekly training volume, and BMI were the main predictors of marathon performance, with higher VO_2max and lower BMI associated with faster performance.	Moderate
Mantzios et al., 2022	Decision Tree	R^2	Air temperature was the most important environmental predictor (40%), with an optimal WBGT of 7.5–15°C; every +1°C decreased performance by 0.3–0.4%.	Excellent

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tomaszewski et al., 2024	Random Forest, XGBoost & LightGBM	R^2, MAPE, RMSE	Machine learning accurately predicted lactate threshold, with LightGBM best for AnT and Random Forest best for AeT.	Moderate
Quittmann et al., 2021	Linear Regression	-	The addition of ċLamax increases the accuracy of the 5 km prediction by 4.4%, as well as the ability to finish kick.	Moderate
Wiecha et al., 2022	XGBoost & Linear Regression	R^2 & RMSE	XGBoost and multivariate regression successfully identified important factors of resilience, and the VAT, VRCP, Vmax models showed stable accuracy for all genders.	Excellent
Feely et al., 2023	-	-	CBR is able to provide personalized training recommendations, effectively increasing the motivation and consistency of beginner runners.	Moderate
Johansson et al., 2023	Linear Regression	R^2, MAE	Machine learning is more accurate than manual methods in predicting finish times, and can identify runners at risk of pacing errors.	Excellent
Figueiredo et al., 2021	Linear Regression	R^2	Critical Speed is the best 5 km predictor ($R^2 = 0.90$), outperforming Vpeak_TF.	Moderate
Rothschild et al., 2024	MARS, SVM, KNN, XGBoost, LightGBM, NNET	RMSE	Machine learning successfully predicted daily recovery and HRV more accurately than baseline, with training load, sleep, and HRV as the primary predictors.	Excellent
Vos et al., 2024	GLM & Random Forest	R^2 & MAE	GLM is best for absolute performance prediction, Random Forest is best for performance change, influenced by VO_2max, ventilatory thresholds, body composition, training impulse, and well-being.	Moderate
Olaya-Cuartero et al., 2023	Linear Regression	R^2	The half-marathon race is a valid determinant of CP, with a strong relationship between CP and race power ($r = 0.88$), making it practical and non-invasive.	Moderate
Pirscoveanu & Oliveira, 2024	Linear Regression, SVN, GPR.	RMSE	The machine learning model accurately predicts RPE using smartwatch data, such as cadence, stride length, heart rate, and oscillation.	Excellent
Lerebourg et al., 2022	AN & KNN	R^2 & MAE	ANN and KNN accurately predicted marathon times, with KNN slightly superior; predictors included 10km time, age, gender, and BMI.	Excellent
Coquart, 2023	Linear Regression	-	Linear Regression predicts 100 km performance with good accuracy ($R^2 = 0.49$) using marathon performance and environmental variables.	Moderate
Alvero-Cruz et al., 2020	-	-	Long-distance performance prediction is primarily influenced by	Moderate

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

			VO_2max and training load, as the most dominant physiological and training factors.	
Muñoz-Pérez et al., 2024	*Multiple Regression* (Tidak Spesifik)	R^2 & RMSE	Pacing strategy and half-marathon performance are strong predictors of marathon performance, especially even/negative pacing.	Excellent
Nikolaidis & Knechtle, 2023	Linear Regression	R^2	Half-marathon times can be accurately predicted from BMI, VO_2max, and weekly training distance.	Moderate
Keogh et al., 2020	*Multiple Regression* (Tidak Spesifik)	R^2	Training history was the most dominant predictor of marathon performance, with the model explaining 85.8% of performance variability.	Excellent
Fokkema et al., 2020	Linear Regression & Logistic Regression	-	High training volume and longer long runs increase finishing speed, without increasing the risk of injury.	Moderate
Knechtle et al., 2024	XGBoost	R^2 & MAE	Race location was the strongest predictor of speed (53%), followed by the athlete's country of origin (21%), indicating the role of culture and geography.	Moderate
Knopp et al., 2025	-	-	Physiological parameters such as VO_2max, running economy, and ventilatory thresholds are essential in predicting elite performance.	Moderate
Dash, 2024	LSTM	Accuracy	The LSTM model outperformed the traditional model with 90.4% accuracy in predicting race distances.	Excellent
Casado et al., 2019	Linear Regression	R^2	Elite runners' performance is heavily influenced by the volume of easy runs, tempo runs, and short intervals, which increase total mileage and long-term aerobic adaptation.	Excellent

Pipeline and Evaluation Model

Based on the results of the systematic review, the pipeline carried out by the model in predicting running performance generally consists of five main stages as shown in Figure 3 below.

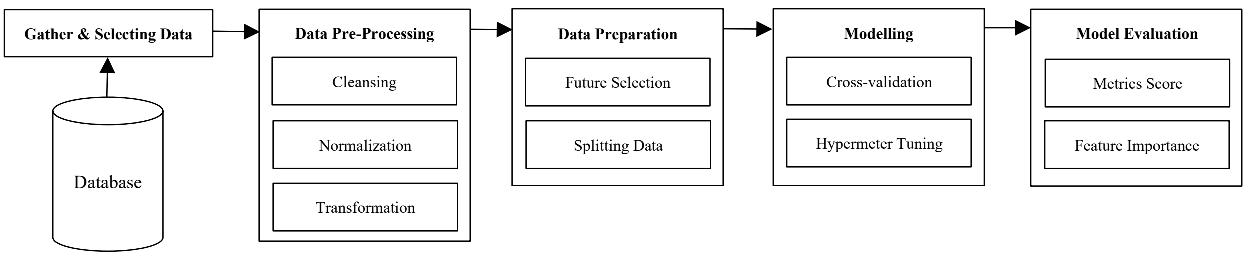

Fig. 3 Machine Learning Pipeline Diagram

Based on the synthesis of 26 studies, the general machine learning pipeline in predicting running performance consists of: (1) selecting data from running activities or physiological/non-physiological tests from the database. (2) Data preparation is carried out in several stages, namely cleansing, normalization, and data transformation. (3) Data pre-processing by performing feature engineering in the form of extracting existing variables into a new variable, for example pace, heart rate (HR), lactate, and so on. The data splitting stage is also carried out to separate

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the model training and test data. (4) The model stage is carried out with the cross-validation stage, hyperparameter tuning and feature selection. (5) The model evaluation stage is carried out to see the metrics score and feature importance for model interpretation so that it is easy to understand.

Most studies use R^2 and RMSE metrics to evaluate the performance of the models used. The ensemble learning model shows an R^2 value above 0.80 and RMSE below 10%. On the other hand, the use of a deep learning model can produce more accurate values when compared to the Linear Regression model with an increase of almost 50% (Zhou, 2025). This aims to avoid overfitting on small datasets, as reported in studies (Dash, 2024a; Vos et al., 2024). Evaluation of model stability shows that cross-validation with Stratified K-Fold and Bayesian optimization for hyperparameter tuning are the most frequently used strategies to avoid overfitting.

Literature Review Summary

The utilization of machine learning in predicting running performance is currently developing rapidly, at least in the last 5 years (Wiecha et al., 2022). The ability of machine learning to identify complex patterns from physiological, biomedical, and environmental data makes this method chosen in predicting running performance compared to conventional methods that tend to be linear and unable to capture complex relationships (Tomaszewski et al., 2024). Accurate prediction capabilities can provide time or speed estimates and training planning that can be useful for athletes and recreational athletes (Nikolaidis & Knechtle, 2023). Based on the results of the literature review, the most frequently used machine learning models in the context of running performance include multivariate Linear Regression, decision tree-based models (Random Forest, XGBoost, LightGBM) and deep learning (LSTM, RNN, GRU). To illustrate the machine learning approaches applied in running performance prediction. Table 3 summarizes the models, input variables, and key findings identified in the reviewed studies.

Table 3. Machine Learning Models Summary

Author and Year	Distance / Context	Machine Learning Model	Main Input Variable	Key Findings
Casado et al., 2019; Fokkema et al., 2020; Nikolaidis et al., 2021	5 km - Marathon	Multi Linear Regression	VO_2max, BMI, age, gender, training frequency	Linear Regression demonstrated moderate predictive performance ($R^2 \approx 0.60–0.77$) with high interpretability but limited ability to model non-linear relationships.
Figueiredo et al., 2021; Johansson et al., 2023	Marathon	Linear Regression	Physiological and anthropometric variables	Regression models were effective for small datasets but showed lower accuracy compared to advanced ML methods.
Mantzios et al., 2022; Tomaszewski et al., 2024; Wiecha et al., 2022	Marathon – Ultramarathon	Random Forest, XGBoost, LigtGBM	Physiological and anthropometric variables	Ensemble learning models improved prediction accuracy by approximately 15–20% compared to Linear Regression.
Rothschild et al., 2024; Vos et al., 2024	Endurance running	XGBoost, Random Forest	Physiological and training-related variables	Ensemble models handled complex physiological data effectively but required explainability techniques to improve interpretability.
Dash, 2024	Multi-distance races	LSTM	Physiological and training-related variables	LSTM generalized performance prediction across distances with accuracy up to 90.4%, outperforming traditional models.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Zhou, 2025	Marathon	GRU	Pace, elevation, historical race data	GRU achieved strong predictive performance (MAPE 5.2%, RMSE 0.11 h) by modeling temporal training dynamics.
Nikolaidis et al., 2021	5 km	Ensemble Learning	VO_2max (~40%), BMI, structured training	VO_2max emerged as the dominant predictor for 5 km performance, while ensemble models showed potential for integrating wearable and training data..

Table 3 summarizes the machine learning models applied in running performance prediction studies, including the modeling approaches used, input variable categories, and reported evaluation metrics across the selected literature.

DISCUSSIONS

A thematic analysis was conducted to identify key patterns and trends across the 26 articles that passed the inclusion stage. The primary objective of this analysis was to identify common research directions and variations in approaches used in the application of machine learning to predict running performance, particularly in the context of the 5 km distance. These four themes jointly address the research objectives of this systematic review by examining (1) how machine learning approaches have been applied to predict running performance, (2) which physiological, anthropometric, and contextual features are most influential, and (3) how predictive models are increasingly personalized, with particular relevance to the 5 km running context. Based on the literature synthesis, four main themes emerged, which complement each other and illustrate the development of research in this field, which are explained below.

Theme 1 - Application of Machine Learning Models in Running Performance Prediction

In the reviewed literature, three dominant modeling paradigms were consistently applied to predict running performance: linear regression-based models, ensemble learning algorithms, and deep learning approaches. These methods differ primarily in their ability to capture nonlinear relationships, temporal dynamics, and model interpretability.

The use of Linear Regression machine learning models is widely applied in studies with relatively limited input datasets, especially on physiological variables such as VO_2max, BMI, age, gender, and training frequency (Casado et al., 2019; Figueiredo et al., 2021; Fokkema et al., 2020; Johansson et al., 2023; Nikolaidis et al., 2021; Quittmann et al., 2021). The choice of the Linear Regression model is often chosen by researchers due to its ease of model interpretation, although with the consequence of relatively small model metric results ($R^2 = 0.60 – 0.77$) (Coquart, 2023; Nikolaidis et al., 2021; Nikolaidis & Knechtle, 2023; Olaya-Cuartero et al., 2023) when compared to ensemble learning models.

In contrast, the use of ensemble machine learning models (XGBoost, Random Forest, LightGBM) shows the ability to handle the complexity of physiological data possessed by runners (Knechtle et al., 2024; Mantzios et al., 2022; Thuany et al., 2025; Tomaszewski et al., 2024; Vos et al., 2024; Wiecha et al., 2022). Most ensemble-based models reported an improvement in predictive, as measured by R^2 or RMSE compared to the Linear Regression model in the context of running performance. This improved performance is largely due to their ability to capture nonlinear relationships and higher-order interactions among physiological, anthropometric, and contextual variables. However, reduced interpretability remains a recurring limitation, which some studies have addressed through post-hoc explanation techniques such as feature importance ranking and SHAP analysis.

Deep learning approaches, including RNN, LSTM, and GRU architectures, represent a more recent trend in running performance prediction. These models are particularly effective in leveraging sequential training data to capture temporal patterns related to fatigue accumulation, pacing behavior, and training adaptation. While deep learning models generally achieve the highest predictive accuracy among the reviewed approaches, their applicability is constrained by substantial data requirements, limited transparency, and increased risk of overfitting when applied to small or distance-specific datasets (Vos et al., 2024).

In the specific context of 5km running, evidence suggests that predictive models rely heavily on dominant physiological features particularly VO_2max, complemented by training structure and body composition variables (Nikolaidis et al., 2021). However, only a small proportion of the reviewed studies explicitly focus on 5km

*name of corresponding author

SinkrOn
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

performance, highlighting a structural gap in the literature. These limitations indicate that, despite methodological advances, the development of robust and generalizable machine learning models for 5 km performance prediction remains limited by data availability rather than algorithmic capabilities.

Key takeaway: Ensemble and deep learning models consistently outperform linear regression in running performance prediction, but ensemble methods currently offer the most practical solution for 5 km applications due to their balance between accuracy, interpretability, and data demands.

Theme 2 - Physiological and Anthropometric Factors as Predictors of Running Performance

In several studies, physiological variables consistently emerged as the feature importance predictors of running performance at distances ranging from 5 km to the marathon. Indicators of aerobic capacity, specifically VO_2max, running efficiency, and lactate threshold, were repeatedly identified as important features in linear regression, ensemble learning, and deep learning models. These findings indicate that aerobic metabolism remains a key determinant of running speed and finish time, regardless of the modeling approach.

Anthropometric factors, such as body mass index (BMI), body fat percentage (BFP), and muscle mass, are also commonly included as predictive features, although their role is primarily modulatory rather than dominant. Several studies report that anthropometric variables improve model stability and individual-level prediction accuracy by accounting for biomechanical efficiency and metabolic load. For example, Nikolaidis et al. (2021) showed that combining VO_2max with BMI and training volume yields significantly higher explanatory power than using a single physiological indicator, highlighting the importance of integrating physiological and body composition measurements.

The relevance of physiological and anthropometric features in the context of nonlinear modeling has been supported by several studies. Using large-scale datasets, Wiecha et al. (2022) and Tomaszewski et al. (2024) showed that variables related to VO_2 at the primary metabolic threshold (e.g., VO_2 at the point of respiratory compensation) consistently ranked as the most influential predictor in XGBoost and Random Forest models. These results demonstrate that machine learning algorithms are capable of capturing the complex interactions between aerobic capacity and physiological characteristics that are not fully represented by traditional linear models.

Importantly, several studies have emphasized the contribution of anaerobic factors, particularly in middle-distance running. (Quittmann et al., 2021) showed that incorporating maximal lactate accumulation rate (ċLamax) significantly improved prediction of 5 km running performance, confirming that anaerobic capacity makes a significant contribution to race outcome, particularly during the final acceleration phase. This finding is particularly relevant for 5-km runners, where tactical speed and the finishing kick play a crucial role.

Overall, the synthesis of evidence indicates that physiological variables explain the majority of performance variance, while anthropometric factors enhance model robustness and personalization. For future machine learning applications, hybrid models that integrate physiological indicators with wearable-derived features (e.g., heart rate variability, cadence, pace dynamics) appear promising for improving generalization and individualized prediction accuracy, particularly in the context of 5 km running performance.

Key takeaway: Across studies, physiological variables especially VO_2max and lactate-related indicators consistently represent the strongest predictors of running performance, while anthropometric factors enhance model stability and personalization, particularly for 5 km runners.

Theme 3 - Non-Physiological and Contextual Factors in Predictive Models

Running performance was also influenced by non-physiological and contextual factors that emerged as important variables. Environmental conditions, pacing strategies, experience, training volume, and geography were repeatedly identified as influential factors across various distances, including the 5km run, marathon, and ultramarathon. These factors explain performance variations that cannot be fully explained by human biological parameters.

Environmental variables, particularly air temperature, humidity, wind speed, and sunlight exposure, are among the most consistently reported contextual predictors. Studies have shown that suboptimal environmental temperature conditions negatively impact running speed and pace stability, with performance sensitivity increasing as race duration progresses (Mantzios et al., 2022). These findings suggest that environmental effects interact with physiological capacity, amplifying fatigue and reducing efficiency, even among runners with comparable aerobic fitness.

Pacing behavior and race strategy also play a significant role in performance outcomes. Evidence from multiple studies suggests that even or negative pacing strategies are associated with faster finish times and reduced performance variability, particularly in longer races (Coquart, 2023; Knechtle et al., 2024; Muñoz-Pérez et al., 2024). Importantly, pace-related features have been shown to retain predictive value even when controlling for physiological variables, highlighting their relevance as independent contextual predictors reflecting tactical decision-making and experience.

*name of corresponding author

SinkrOn
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

A runner's training history and experience further contribute to performance prediction. Studies consistently report that accumulated training volume, prior race exposure, and realistic goal setting are associated with improved pace control and more stable race execution (Keogh et al., 2020). These findings underscore the importance of incorporating longitudinal training context into predictive models, rather than relying solely on momentary physiological assessments.

Geographic and locational factors, including race location and athlete origin, have also been identified as influential predictors in several studies (Knechtle et al., 2024). While these factors likely act as proxies for climate, altitude, and training culture, their inclusion in improved model performance on large-scale datasets suggests that regional context can meaningfully shape running outcomes.

The findings from these studies suggest that non-physiological factors can improve the performance of machine learning models by incorporating real-world race conditions into performance predictions. For the 5km run, where physiological capacity is relatively homogeneous among competitive runners, contextual variables such as temperature, pacing behavior, and recent training load can play a disproportionate role in differentiating performance outcomes. Therefore, integrating contextual and physiological data is a crucial step towards more accurate and personalized running performance predictions.

Overall, non-physiological and contextual factors play a significant role in improving the performance of predictive models, particularly over short-to-medium distances such as the 5km. Incorporating environmental data, pacing strategies, and geographic information into machine learning models broadens the scope of predictive performance, representing an important step towards more accurate and personalized running performance predictions.

Key takeaway: Non-physiological contextual factors particularly environmental conditions, pacing strategies, training history, and geographic context contribute to running performance variability and enhance the real-world applicability of predictive models, especially for differentiating outcomes in 5 km events.

Theme 4 - Personalized Athlete Training and Monitoring Based on Machine Learning

Across the studies reviewed, machine learning-based personalization consistently emerged as a key advancement for improving performance prediction and training adaptation in endurance running. Unlike traditional, one-size-fits-all training approaches, personalized models leverage individual physiological profiles, training history, and contextual data to account for interindividual variability in performance responses and recovery dynamics.

Sensor data collected by runners plays a crucial role in enabling personalization by providing continuous and individualized measurements of training behavior and physiological responses. Across studies, features derived from heart rate, pace, cadence or stride, and sleep patterns have repeatedly been identified as key inputs for individualized predictions (Pirscoveanu & Oliveira, 2024; Rothschild et al., 2024). Machine learning approaches (including ensemble learning , deep learning, and case-based reasoning) demonstrate the ability to integrate data streams into adaptive models that evolve with athletes over time.

Other studies highlight the potential of personalized models to inform training load adjustments, optimize recovery strategies, and reduce injury risk by detecting early signs of excessive fatigue or maladaptation (Feely et al., 2023; Zhou, 2025). These applications underscore the role of machine learning as a tool for dynamic training management, rather than static outcome prediction.

In the context of 5km running, machine learning-based personalization is particularly relevant due to the high training intensity and the strong influence of short-term fatigue and recovery on race-day performance. Given the relatively narrow physiological performance range among competitive 5km runners, individual responses to training stimuli and daily preparedness can play a significant role in performance outcomes. Therefore, future research should prioritize the development of hybrid personalization workflows that integrate physiological indicators, anthropometric characteristics, and behavioral features derived from wearable devices into an integrated, end-to-end training prediction and support system.

Key takeaway: Machine learning-based personalization enhances running performance prediction and training adaptation by capturing individual-specific physiological and behavioral responses, with particularly strong relevance for 5 km runners where short-term fatigue and recovery dynamics critically influence performance.

CONCLUSION

Based on the analysis using the PRISMA method on 26 studies that met the inclusion criteria, it can be concluded that research on machine learning-based running performance prediction shows significant progress, both methodologically and applicably. The four main themes discussed in this review complement each other and provide a comprehensive understanding of how machine learning can be applied to understand, predict, and optimize runner performance in various contexts, including the 5 km distance.

The first theme demonstrates that machine learning models, particularly ensemble learning (XGBoost, Random Forest, and LightGBM) and deep learning (LSTM and GRU), have advantages in capturing non-linear

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15743
e-ISSN : 2541-2019
p-ISSN : 2541-044X

relationships and temporal patterns from physiological data and repetitive training rather than conventional linear regression. This results in higher prediction accuracy than conventional Linear Regression models, although they still face challenges in interpretation and require large data sets.

The second theme confirms that physiological factors such as VO_2max, running economy, and lactate threshold remain key predictors of running performance, especially over medium distances like the 5 km. Anthropometric factors such as BMI and body fat percentage also demonstrate a significant influence on runners' work efficiency and metabolic load. The combination of these two factors has been shown to provide more precise and stable predictive models.

The third theme broadens understanding by highlighting the influence of non-physiological factors such as environmental conditions (temperature, humidity, wind speed), pacing strategy, race experience, and geographic variables. These factors, while not directly related to physiological capacity, contribute significantly to performance variation and improve model prediction precision when interacted with physiological variables.

The fourth theme marks a paradigm shift toward personalized data-driven training and monitoring (personalized machine learning systems). By integrating data from wearable sensors, HRV, and training behavior, models (GRU, LightGBM), and Case-Based Reasoning enable training systems that adapt to an individual's physiological and psychological state in real time.

This review is, to our knowledge, the first to systematically synthesize machine learning algorithms, feature sets, and personalization strategies for running performance prediction across distances, with particular emphasis on 5 km events. By integrating methodological, physiological, contextual, and personalization perspectives, this study provides a structured foundation for future research and practical applications in sport analytics.

Several limitations of this review should be acknowledged. First, the literature search was restricted to selected academic databases and included only peer-reviewed articles published in English, which may have resulted in the exclusion of relevant studies from other databases or non-English sources. Second, the reviewed literature is predominantly composed of studies focusing on marathon and ultramarathon events, leading to an imbalance in distance-specific evidence and a relatively limited number of studies dedicated exclusively to 5 km performance prediction. Third, this review did not conduct a quantitative meta-analysis due to substantial heterogeneity in study designs, datasets, input features, outcome variables, and evaluation metrics, which limited direct statistical comparison across studies.

REFERENCES

Alvero-Cruz, J. R., Carnero, E. A., García, M. A. G., Cárceles, F. A., Correas-Gómez, L., Rosemann, T., Nikolaidis, P. T., & Knechtle, B. (2020). Predictive performance models in long-distance runners: A narrative review. In *International Journal of Environmental Research and Public Health* (Vol. 17, Issue 21, pp. 1–22). MDPI. https://doi.org/10.3390/ijerph17218289

Azzahra, F. S. P., & Prianto, C. (2025). Tinjauan Literatur Sistematis: Analisis Implementasi Kecerdasan Buatan untuk Verifikasi Dokumen. *Jurnal Ilmiah Sistem Informasi*, *4*(3), 417–430. https://doi.org/10.51903/kjjwk708

Casado, A., Hanley, B., Santos-Concejero, J., & Ruiz-P ´ Erez, L. M. (2019). *World-Class Long-Distance Running Performances Are Best Predicted by Volume of Easy Runs and Deliberate Practice of Short-Interval and Tempo Runs*. www.nsca.com

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, *13*(1). https://doi.org/10.1186/s12916-014-0241-z

Coquart, J. B. (2023). Prediction of performance in a 100-km run from a simple equation. *PLoS ONE*, *18*(3 March). https://doi.org/10.1371/journal.pone.0279662

Dash, S. (2024a). Win Your Race Goal: A Generalized Approach to Prediction of Running Performance. *Sports Medicine International Open*, *08*(CP). https://doi.org/10.1055/a-2401-6234

Dash, S. (2024b). Win Your Race Goal: A Generalized Approach to Prediction of Running Performance. *Sports Medicine International Open*, *08*(CP). https://doi.org/10.1055/a-2401-6234

Feely, C., Caulfield, B., Lawlor, A., & Smyth, B. (2023). Modelling the Training Practices of Recreational Marathon Runners to Make Personalised Training Recommendations. *UMAP 2023 - Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 183–193. https://doi.org/10.1145/3565472.3592952

Figueiredo, D. H., Figueiredo, D. H., Manoel, F. de A., & Machado, F. A. (2021). Peak Running Velocity or Critical Speed Under Field Conditions: Which Best Predicts 5-km Running Performance in Recreational Runners? *Frontiers in Physiology*, *12*. https://doi.org/10.3389/fphys.2021.680790

*name of corresponding author

Fokkema, T., van Damme, A. A. D. N., Fornerod, M. W. J., de Vos, R. J., Bierma-Zeinstra, S. M. A., & van Middelkoop, M. (2020). Training for a (half-)marathon: Training volume and longest endurance run related to performance and running injuries. *Scandinavian Journal of Medicine and Science in Sports*, *30*(9), 1692–1704. https://doi.org/10.1111/sms.13725

Garmin Indonesia. (2024). *Garmin Ungkap Tren Kebugaran di Tahun 2024: Lari, Tenis dan Golf Jadi Olahraga Paling Digemari di Indonesia*. Https://Www.Garmin.Co.Id/News/Press-Release/News-2024-Dec-Garmin-Connect-Data-Report/.

Indrisari, E., Febiansyah, H., & Adiwinoto, B. (2025). A Systematic Literature Review on the Application of Machine Learning for Predicting Stunting Prevalence in Indonesia (2020–2024). *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *14*(3), 277–283. https://doi.org/10.32736/sisfokom.v14i3.2366

Johansson, M., Atterfors, J., & Lamm, J. (2023). Pacing Patterns of Half-Marathon Runners: An analysis of ten years of results from Gothenburg Half Marathon. *International Journal of Computer Science in Sport*, *22*(1), 124–138. https://doi.org/10.2478/ijcss-2023-0014

Keogh, A., O'connor Sheridan, O., Mccaffrey, O., Dunne, S., Lally, A., & Doherty, C. (2020). *The Determinants of Marathon Performance: An Observational Analysis of Anthropometric, Pre-race and In-race Variables*. http://www.intjexersci.com

Knechtle, B., Weiss, K., Valero, D., Villiger, E., Nikolaidis, P. T., Andrade, M. S., Scheer, V., Cuk, I., Gajda, R., & Thuany, M. (2024). Using machine learning to determine the nationalities of the fastest 100-mile ultra-marathoners and identify top racing events. *PLoS ONE*, *19*(8). https://doi.org/10.1371/journal.pone.0303960

Knopp, M., Guppy, F., Joyner, M., Muniz-Pardos, B., Wackerhage, H., Schönfelder, M., Pitsiladis, Y., & Ruiz, D. (2025). Validation of marathon performance model based on physiological factors in world-class East African runners: a case series. *Translational Exercise Biomedicine*, *2*(1), 1–8. https://doi.org/10.1515/teb-2024-0016

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

Mantzios, K., Ioannou, L. G., Panagiotaki, Z. O. E., Ziaka, S., Périard, J. D., Racinais, S., Nybo, L., & Flouris, A. D. (2022). Effects of Weather Parameters on Endurance Running Performance: Discipline-specific Analysis of 1258 Races. *Medicine and Science in Sports and Exercise*, *54*(1), 153–161. https://doi.org/10.1249/MSS.0000000000002769

Muñoz-Pérez, I., Castañeda-Babarro, A., Santisteban, A., & Varela-Sanz, A. (2024). Predictive performance models in marathon based on half-marathon, age group and pacing behavior. *Sport Sciences for Health*, *20*(3), 797–810. https://doi.org/10.1007/s11332-023-01159-4

Nikolaidis, P. T., & Knechtle, B. (2023). Predictors of Half-Marathon Performance in Male Recreational Athletes. *EXCLI Journal*, *22*, 559–566. https://doi.org/10.17179/excli2023-6198

Nikolaidis, P. T., Rosemann, T., & Knechtle, B. (2021). Development and Validation of Prediction Equation of "Athens Authentic Marathon" Men's Race Speed. *Frontiers in Physiology*, *12*. https://doi.org/10.3389/fphys.2021.682359

Olaya-Cuartero, J., Pueo, B., Villalon-Gasch, L., & Jiménez-Olmedo, J. M. (2023). Prediction of Half-Marathon Power Target using the 9/3-Minute Running Critical Power Test. *Journal of Sports Science and Medicine*, *22*(3), 525–530. https://doi.org/10.52082/jssm.2023.525

Pirscoveanu, C. I., & Oliveira, A. S. (2024). Prediction of instantaneous perceived effort during outdoor running using accelerometry and machine learning. *European Journal of Applied Physiology*, *124*(3), 963–973. https://doi.org/10.1007/s00421-023-05322-0

Quittmann, O. J., Foitschik, T., Vafa, R., Freitag, F. J., Sparmann, N., Nolte, S., & Abel, T. (2021). Is Maximal Lactate Accumulation Rate Promising for Improving 5000-m Prediction in Running? *International Journal of Sports Medicine*, *44*(4), 268–279. https://doi.org/10.1055/a-1958-3876

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

Thuany, M., Weiss, K., Valero, D., Villiger, E., Andrade, M. S., Nikolaidis, P. T., Scheer, V., de Lira, C. A. B., Vancini, R. L., Cuk, I., Braschler, L., Rosemann, T., & Knechtle, B. (2025). An analysis of the 6-h ultra-marathon race using a machine learning approach. *Frontiers in Sports and Active Living*, *7*. https://doi.org/10.3389/fspor.2025.1577470

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Tomaszewski, M., Lukanova-Jakubowska, A., Majorczyk, E., & Dzierżanowski, Ł. (2024). From data to decision: Machine learning determination of aerobic and anaerobic thresholds in athletes. *PLoS ONE*, *19*(8). https://doi.org/10.1371/journal.pone.0309427

Vijay, S. A., Sivakumar, C., Kumar, P. V., Muralidharan, C. K., Rajkumar, K. V., Kannan, K. R., Pradeepa, M., Sivasankar, P., Mariam, A. A., & Anand, U. K. A. (2024). Lactate Threshold Training to Improve Long-Distance Running Performance: A Narrative Review. *Montenegrin Journal of Sports Science and Medicine*, *20*(1), 19–29. https://doi.org/10.26773/mjssm.240303

Vos, L., Vergeer, R., Goulding, R., Weide, G., de Koning, J., Jaspers, R., & Zwaard, S. van der. (2024). *Predicting physical performance after training: insights from machine learning using small samples*. https://doi.org/10.21203/rs.3.rs-4707433/v1

Weiss, K., Valero, D., Villiger, E., Scheer, V., Thuany, M., Aidar, F. J., de Souza, R. F., Cuk, I., Nikolaidis, P. T., Rosemann, T., & Knechtle, B. (2024). Associations between environmental factors and running performance: An observational study of the Berlin Marathon. *PLoS ONE*, *19*(10). https://doi.org/10.1371/journal.pone.0312097

Westwood, M., Deeks, J. J., Whiting, P. F., Weswood, M. E., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. N., & Kleijnen, J. (2011). QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *Article in Annals of Internal Medicine*. https://doi.org/10.1059/0003-4819-155-8-201110180-00009

Wiecha, S., Kasiak, P. S., Cieśliński, I., Maciejczyk, M., Mamcarz, A., & Śliż, D. (2022). Modeling Physiological Predictors of Running Velocity for Endurance Athletes. *Journal of Clinical Medicine*, *11*(22). https://doi.org/10.3390/jcm11226688

Winandy Soesilo, F., & Suprapti Hendharta, C. (2025). Factors Affecting Runners' Continuance Intention on Strava Running Apps Usage In Jakarta. In *Jurnal Bisnis & Komunikasi* (Vol. 12, Issue 3). https://doi.org/https://ojs.kalbis.ac.id/index.php/kalbisocio/en/article/download/4574/1161/14635

Zhou, S. (2025). DeepRace: GRU-Based Sequence Modeling Framework for Marathon Performance Prediction. *Informatica (Slovenia)*, *49*(28), 153–160. https://doi.org/10.31449/inf.v49i28.8818

==

*name of corresponding author