

Predicting Elite 100m Sprint Performance: A Comparative Analysis of Regression and Machine Learning Model

Wardany Kusumasari^{1*}, Tabah Dimas Pujawardana², Sandy Vantika³, and
Yusri Kamilatul Huda⁴

^{1,3,4}Department of Mathematics, Bandung Institute of Technology, Bandung, Indonesia

²Department of Physical Education, Jenderal Soedirman University, Purwokerto, Indonesia

*Corresponding author: wardanykusumasuwardi@gmail.com

ABSTRACT

Sprint performance can be analyzed through a data-driven approach by utilizing biomechanical parameters that play an important role in athletic development. The two biomechanical parameters explored in this study are Cadence and step length which are analyzed for their influence on running speed. Data was taken from six elite world sprinters in the men's 100 meter final of the 2020 Tokyo Olympics focusing on performance dynamics at every 10 meter interval to capture the real performance of athletes in competition. The modeling approach uses regression and machine learning, namely Random Forest (RF) and Neural Network (NN). Linear regression is used as a basic model that is easy to interpret, while Random Forest and Neural Network are used in the digital-based automated performance analysis system. These three models will be evaluated using the coefficient of determination (R^2) and error (MSE) to compare their accuracy. The results show that all three models are capable of providing high accuracy in predicting running speed. The regression model provides strong statistical validity with R^2 of 0.9217 and MSE of 0.0062. The Random Forest model produced the highest prediction accuracy with R^2 of 0.9757 and MSE of 0.1027. Meanwhile, the Neural Network offers a more flexible approach in capturing complex patterns from the data with R^2 of 0.9261 and MSE of 0.3117. These results confirm the importance of integrating digital technology and intelligent predictive models in optimizing athletic performance. Thus, training programs can be made more specific and data-driven, improving the efficiency and effectiveness of athlete performance development.

Keywords: Cadence, Machine learning, Regression, Sprint velocity prediction, Step length.

1. INTRODUCTION

Advances in sports science and biomechanics have opened up new insights into the factors that influence athlete performance in international competitions such as the Olympics. For example, in 100-meter sprint competition, a runner's speed is not solely dependent on muscle strength, but is also the result of a complex interaction between a number of biomechanical variables and other external factors (Mero, Komi, & Gregor, 1992).

In an effort to improve running speed, coaches and athletes usually focus on two main biomechanical components, namely cadence and step length. These two components are usually developed through structured training programs. Achieving an optimal balance between cadence and step length is a challenge in the quest for maximum speed.

Research by Sobarna et al. (2023) indicated that training targeting an increase in step length had a greater effect on sprint performance compared to training to increase cadence. This suggests that while both can improve speed, their effectiveness depends on which

biomechanical aspect is improved. However, adaptation to step length training tends to take longer than cadence training.

According to Hunter et al. (2004), there is a mutually influential relationship between step length and cadence; runners with longer steps generally have lower frequencies, and vice versa. This interaction emphasizes the importance of striking a balance between cadence and step length. Training that overemphasizes either aspect may decrease overall sprint efficiency.

Supratman et al. (2024) also stated that training to increase cadence and step length has a significant contribution to increasing average speed, number of steps per second, and leg muscle strength. The consistency of performance results over time suggests that the combination of the two training methods can proportionally produce more optimal performance progress.

In athletic training practice, the variables of cadence and step length are indeed key elements in the development of athletes' training techniques and strategies (Ruiter & Dieen, 2019). Therefore, understanding the relationship between these two variables is crucial to designing an effective training program to improve sprinting ability.

Research conducted by Hussein (2024) also highlighted the importance of biomechanical variables in supporting running performance. The results stated that speed in the first 10 meters, which is influenced by cadence, is the main biomechanical factor that determines the performance of elite runners.

Barbosa et al. (2016) stated that the difference in performance between international and national elite runners in the 100-meter sprint was due to a number of critical moments during the race related to biomechanical factors. Similarly, research by Valunpion & Rangubhet (2025) found that the pattern and dynamics of runners' performance in international events are influenced by preferences for certain techniques, such as the Banzai start technique (37.8%), shoulder-width hand position (51.1%), and hip elevation higher than the shoulder (64.4%).

In the field of sports performance analysis and prediction, mathematical models play an important role. A model is defined as a representation of an entity or system, not to depict its entire complexity, but rather important relevant aspects. Mathematical models represent phenomena from non-mathematical domains through mathematical symbols and relationships. The process of building these models, known as mathematical modeling, involves selecting important elements and making decisions. Through modeling, researchers can understand the fundamental structure and mechanisms of a phenomenon, as well as predict how the phenomenon will evolve under different conditions (Niss, Mogens, & Blum, 2020).

Furthermore, the integration of AI with multivariate time series data from IMUs proves to be a highly effective approach in the field of sports performance analysis. The developed model provides accurate predictions of both fatigue and stamina, allowing training regimens to be personalized and dynamically adjusted. These findings underscore the transformative role AI can play in advancing sports science, signaling a future where data-driven approaches significantly improve athletic training and performance management (Biró, Cuesta-Vargas, & Szilágyi, 2023).

In recent years, in addition to regression, machine learning approaches such as Random Forest (RF) and Neural Network (NN) have been increasingly used for descriptive modeling. For example, the model developed by Agung Prabowo and Ngadiman (2022) was used to predict athletes' travel times in the next Olympics based on historical data. Similarly, the model from Efiong, Olajubu, and Aranuwa (2019) aimed to estimate the fastest times of 100-meter runners in five Olympic seasons.

Based on this description, this study aims to analyze the speed of male runners at the 2018 Olympics based on the variables of cadence and step length, in order to understand how the interaction between the two affects optimal sprint speed. Speed performance models will be analyzed using regression, Random Forest, and Neural Network approaches, and compared based on R^2 and MSE values. The results of this study are expected to be valuable input for coaches and athletes in designing more effective training strategies to improve sprint performance.

2. METHODS

2.1. Data Sources and Subjects

This study used biomechanical data from six elite world sprinters who participated in the men's 100-meter final at the 2020 Tokyo Olympics. Official videos of the games were obtained from the official Olympic channel on YouTube and used as the main source for motion data extraction.

2.2. Data Collection Technique

Data was collected through video analysis using Kinovea software, which enables precise two-dimensional motion tracking on a frame-by-frame basis. Utilizing track markings and travel time, three key biomechanical parameters were measured: cadence (frequency of steps per minute), step length (in meters), and speed (in meters per second). Measurements were taken every 10 meters, from the start line to the finish line, to generate data describing performance dynamics along the 100-meter course.

2.3. Data Analysis Procedure

The measured data were compiled in spreadsheet and analyzed using RStudio software. The analysis was carried out in two main stages:

- Descriptive Analysis

Used to evaluate the development patterns of cadence, step length, and speed along the track. Statistics such as mean and standard deviation were calculated for each 10-meter segment.

- Predictive Modeling

To evaluate the relationship between cadence and step length on running speed, this study uses Linear Regression, Random Forest, and Neural Network based on their distinct advantages in predictive modeling. Linear Regression offers interpretability and statistical clarity (James

et al., 2013), and use of machine learning models, including Random Forest and Neural Networks, is supported by prior research that highlights their relevance for sprint performance prediction using biomechanical inputs such as cadence and step length (Tam and Yao, 2023).

- o Multiple Linear Regression

This approach is used to estimate the linear effect of two predictors on speed. Classical assumption tests will be conducted to ensure model validity. To verify the accuracy of the linear regression model, standard assumption tests were performed, including the Shapiro-Wilk test for normality, the Breusch-Pagan test for homoscedasticity, and the Variance Inflation Factor (VIF) to check for multicollinearity.

- o Random Forest (RF)

This model is used to capture possible non-linear relationships between variables. Parameter tuning and model validation are performed using the *randomForest* library in RStudio. The Random Forest model was developed using 500 decision trees ($n_{tree} = 500$) and was trained in-sample, with default settings used for both $mtry$ and $nodesize$.

- o Neural Networks (NN)

This approach uses an artificial neural network architecture with two inputs (cadence and step length), one or more hidden layers, and one output (speed). The model is trained using the backpropagation method through the *neuralnet* library in RStudio. The Neural Network model was built with a single hidden layer consisting of five neurons. A sigmoid activation function was used in the hidden layer, while a linear activation function was employed in the output layer by setting $linout = TRUE$. The training process involved using the backpropagation algorithm with the BFGS optimization method, and both input and output data were normalized to improve training consistency.

- Model Evaluation

The performance of each model was evaluated using two main metrics:

- o Mean Squared Error (MSE) as an indicator of prediction error, and
- o Coefficient of Determination (R^2) as an indicator of the strength of the relationship between predictors and target variable.

This metric is used as the basis for comparing the accuracy and effectiveness of each modeling approach.

3. FINDING AND DISCUSSION

3.1. Finding

3.1.1. Descriptive statistics

Table 1. Descriptive statistics per 10 meters

Distance (m)	Mean			Deviation Standard		
	Cadence	Step Length	Speed	Cadence	Step Length	Speed
10	226	1,45	5,34	0,0828	4,42	0,0683
20	283	2,06	9,44	0,197	18,5	0,157
30	296	2,17	10	0,358	11,2	0,124
40	295	2,24	10,7	0,466	12,7	0,147
50	294	2,32	11,1	0,396	14,7	0,111
60	293	2,36	11,7	0,461	14,3	0,103
70	292	2,44	12,2	0,329	8,22	0,0752
80	293	2,5	12,1	0,435	6,75	0,0489
90	287	2,57	12,5	0,0612	18,7	0,112
100	282	2,62	12,1	0,263	11,4	0,0649

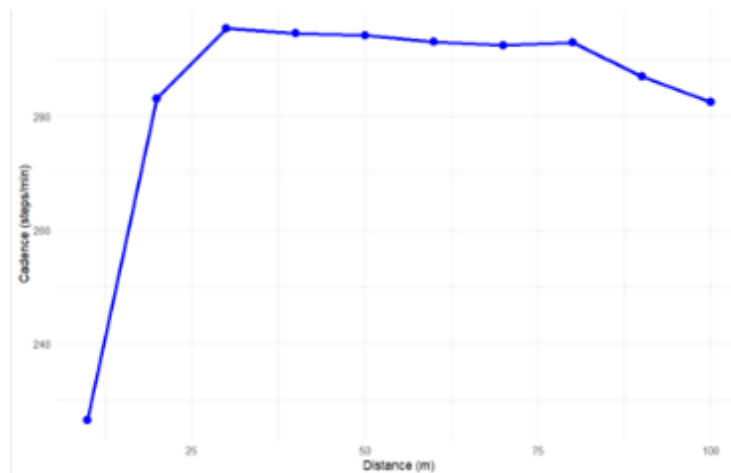


Figure 1. Mean Cadence per 10 Meters

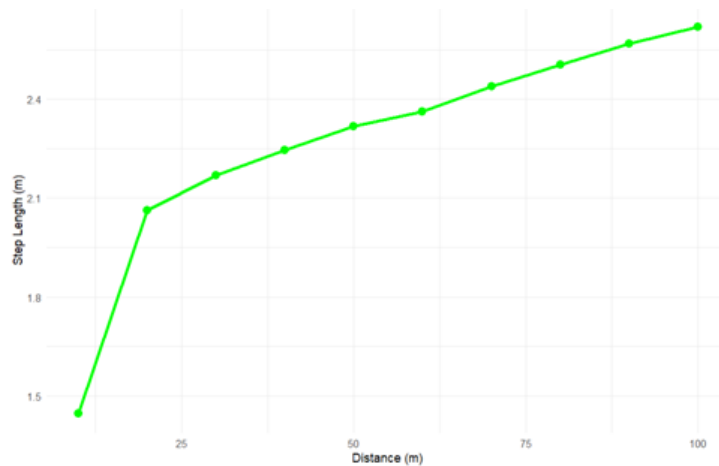


Figure 2. Mean Step Length per 10 Meters

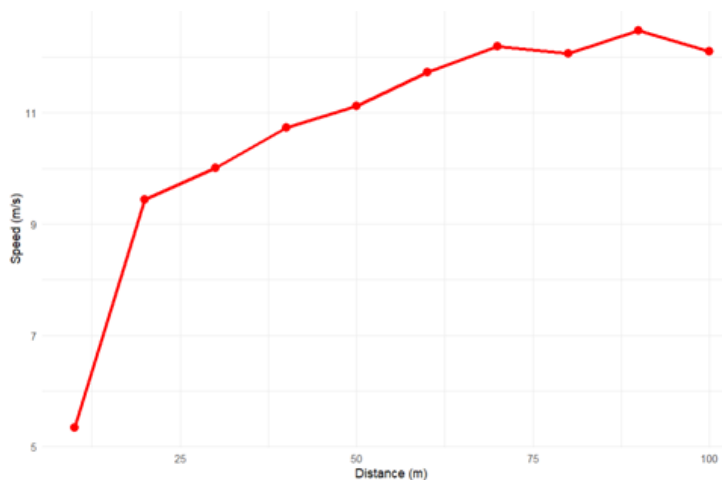


Figure 3. Mean Speed per 10 Meters

3.1.2. Performance Dynamics Throughout the 100-Meter Track

Biomechanical data collected from six elite world runners (Simbine, Jacobs, Kerley, Bingtian, Baker, and Grasse) were analyzed to explore the development of sprint performance along the 100-meter track. Variables studied included cadence (in steps per minute), step length (in meters), and running speed (in m/s), with measurements taken every 10-meter interval.

Descriptive analysis revealed a consistent pattern in speed development: athletes generally experienced an increase in speed from the start line to the 70-90 meter segment, then a slight deceleration towards the finish line. This pattern aligned with the previously recognized phases of sprinting (acceleration, transition, and maximal speed) ending with a mild deceleration phase possibly caused by neuromuscular fatigue.

Step length tends to increase progressively along the course, while step frequency is relatively stable or slightly decreases after mid-race. These two parameters, taken together, proved to be the main determinants of running speed, confirming their relevance in predictive modeling and performance diagnostics.

3.1.3. Predictive Modeling of Running Speed

To quantitatively model the relationship between cadence, step length, and running speed, three predictive modeling approaches were used: multiple linear regression, random forest (RF), and neural networks (NN). The performance of each model was evaluated based on the Mean Squared Error (MSE) and the coefficient of determination (R^2) value to assess the accuracy and influence cadence and step length on speed of each model.

3.1.4. Multiple Linear Regression

Multiple linear regression model with regression equation:

$$\hat{Y} = -0.91286 + 0.31606X_1 + 0.63853X_2$$

The model produced an MSE of 0.0062 and R^2 of 0.9217, indicating that approximately 92.17% of the variance in running speed can be explained by cadence and step length. Assumption testing showed that linearity, homoscedasticity, and residual normality, making the model statistically robust. However, as a linear model, its ability to capture non-linear interactions prevalent in biomechanical performance data is limited.

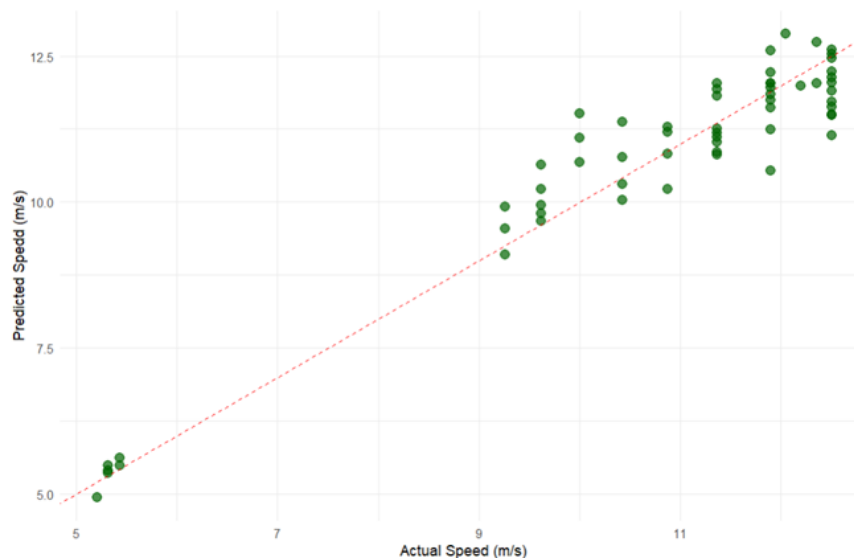


Figure 4. Actual and Predicted speed (Linear Regression model)

3.1.5. Random Forest (RF)

The RF model produced a higher MSE than the linear regression model, which is 0.1027, but with a much higher R^2 value of 0.9757. This suggests that the model was able to explain a larger portion of variance in running speed, potentially due to its ability to capture complex,

non-linear relationships between variables. Feature importance analysis from the Random Forest model indicates that step length (IncNodePurity = 133.36) has a slightly greater influence on sprint speed prediction than cadence (IncNodePurity = 108.87). This suggests that step length contributes more to the model's ability to reduce prediction error.

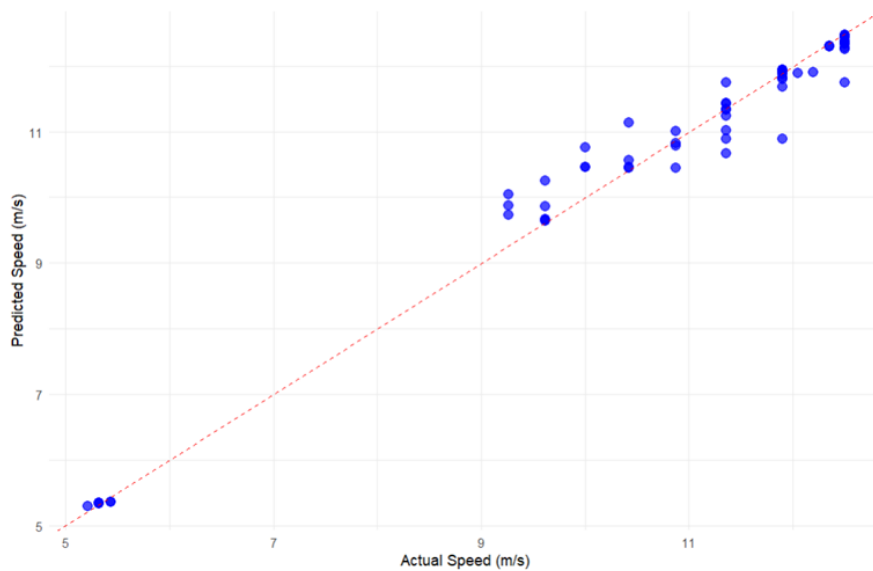


Figure 5. Actual and Predicted speed (Random Forest model)

3.1.6. Neural Network (NN)

A neural network structure that model the relationship between cadence and step length to speed.

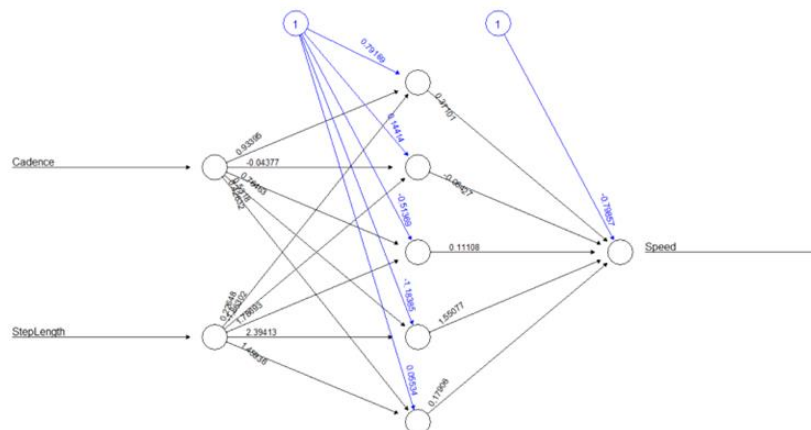


Figure 6. Visualization of Neural Network Architecture

The neural network model showed prediction performance with a significantly higher MSE compared to multiple linear regression and random forest at 0.3117 and an R^2 similar to the regression model at 0.9261. The architecture of these models allows exploration of hidden

features and patterns that often cannot be accommodated by conventional linear approaches. As shown in Figure 6, the architecture consists of a single hidden layer with five neurons, selected based on prior studies that support the adequacy of simple configurations for small datasets and moderately complex relationships (Heaton, 2008; Jain, Mao, & Mohiuddin, 1996). The weights along each connection represent the strength and direction of influence among the nodes. This visualization confirms that the neural network is capable of capturing the complex, non-linear relationships between key biomechanical variables in sprint performance.

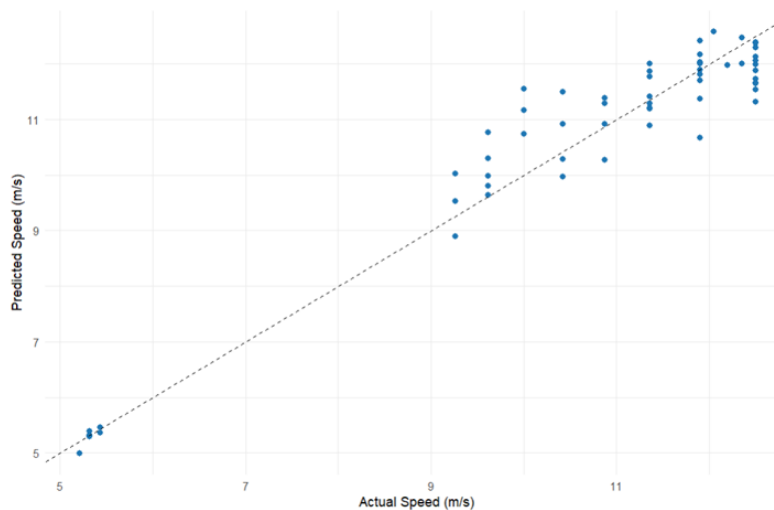


Figure 7. Actual and Predicted speed (Neural Network model)

3.2. Discussion

Analysis of three statistical modeling and machine learning approaches—multiple linear regression, random forest, and neural networks—showed complementary results in understanding and predicting running speed based on key biomechanical parameters: cadence and step length.

The multiple linear regression model offers high interpretability and strong statistical validity. With an R^2 of 0.9217 and an MSE of 0.0062, the model showed that approximately 92% of the variability in speed could be linearly explained by the two predictors. The significant contribution of each variable, with positive coefficients for cadence and step length, confirms the importance of increasing step frequency and step length as strategies for optimizing running speed. As the model also meets the classical assumptions of regression, it can be widely used for athlete performance evaluation in training environments that rely on clean and controlled data.

The Random Forest (RF) model produced the highest prediction accuracy with an R^2 of 0.9757, demonstrating its ability to capture the complexity of non-linear relationships between variables. Despite its slightly higher MSE (0.1027), the model is much more robust in the face of real-world data variability. The implication is that RF is very suitable in the context of real-time monitoring of athlete performance, for example, through wearable sensors or biomechanical tracking devices that generate large and heterogeneous data. RF's ability to

identify non-linear patterns also opens up opportunities to develop personalized training programs that are adaptive to the unique characteristics of each runner.

Meanwhile, neural networks (NN) offer a more flexible approach in extracting complex patterns from data. With an R^2 of 0.9261 and MSE of 0.3117, it performs slightly below RF but shows great potential, especially with large datasets and richer input variables. The network architecture that models layered relationships between variables allows the exploration of hidden features that conventional approaches often cannot accommodate. Practically speaking, the use of NN is promising in the development of artificial intelligence-based automated prediction systems, such as smart coaching assistants that can provide instant feedback to runners.

This study's models, especially Random Forest and Neural Networks, can be applied in real training environments using wearable sensors. Gurchiek et al. (2019) demonstrated that a single accelerometer combined with machine learning could accurately estimate sprint metrics like velocity and acceleration. This supports the use of predictive models in mobile systems or national training programs to provide real-time feedback and improve athlete performance.

Overall, the linear regression approach is suitable for initial interpretation and planning of simple data-driven programs; Random Forest excels in accuracy and is suitable for dynamic data-driven decision-making; while neural networks have great potential in the application of intelligent prediction systems based on digital technology. The combination of the three can be used to build an adaptive, precise, and personalized data-driven athletic training ecosystem.

4. CONCLUSION

This study shows that cadence and step length are key determinants of sprinting speed, with consistent biomechanical patterns observed throughout the 100-meter distance. Multiple Linear Regression offers strong statistical interpretability and reliability ($R^2 = 0.9217$), while Random Forest achieves the highest predictive accuracy ($R^2 = 0.9757$), making it suitable for dynamic and real-time applications. Neural networks, with their flexibility in modeling complex patterns ($R^2 = 0.9261$), show promise for AI-based performance systems. Together, these models provide complementary strengths for developing precise, adaptive and personalized sprint training strategies.

REFERENCES

- Sobarna, A., Hambali, S., Paramitha, S. T., Shafie, M. S., & Ramadhan, M. G. (2023). The Effect of Training Stride Length and Stride Frequency on Increasing Sprint Speed. *Migration Letters*, 20(6), 1122-1136.
- Barbosa, T. M., Forte, P., Marindo. D. A., & Reis, V. M. (2016). Comparison of The World and European Records in The 100m Dash by a Quasi- Physical Model. *Procedia Engineering*, 147, 122-126.
- Biró, A., Cuesta-Vargas, A., & Szilágyi, L. (2023). AI-Assisted Fatigue and Stamina Control for Performance Sports on IMU-Generated Multivariate Times Series Datasets. *Sensors (Basel, Switzerland)*, 24.

- Debaere, S., Jonkers, I., & Delecluse, C. (2013). The Contribution of Step Characteristics to Sprint Running Performance in High-Level Male and Female Athletes. *Journal of Strength and Conditioning Research*, 27(1).
- Adams D., Pozzi, F., Willy, R. W., Carrol, A., Zeni, J. (2018). Altering Cadence Or Vertical Oscillation During Running: Effects On Running Related Injury Factors. *The International Journal of Sports Physical*, 13(4).
- Gurchiek, R. D., Rupasinghe Arachchige Don, H. S., Pelawa Watagoda, L. R., McGinnis, R. S., van Werkhoven, H., Needle, A. R., McBride, J. M., & Arnholt, A. T. (2019). Sprint Assessment Using Machine Learning and a Wearable Accelerometer. *Journal of Applied Biomechanics*, 35(2), 164-169.
- Heaton, J. (2008). *Introduction to neural networks for Java* (2nd ed.). Heaton Research.
- Hussein, E. S. (2024). The Contribution Percentage of Some Biomechanical Variables to The First 10 Meters in The Achievement of Men's 100-Meter Sprint. *International Journal of Disabilities Sports and Health Sciences*, 206-216.
- Hunter, J. P., Marshall, R. N., & McNair, P. J. (2004). Negative Interaction between Step Length and Step Rate Refers to An Increase in One Factor Resulting in A Decrease in The Other. *Med. Sci. Sports Exerc*, 36(2).
- Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, 29(3), 31–44.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
- Kunz, H., & Kaufmann, D. (1981). Biomechanical analysis of sprinting: decathletes versus champions.. *British Journal of Sports Medicine*, 15, 177 - 181.
- Mero, A., Komi, P., & Gregor, R. (1992). Biomechanics of Sprint Running. *Sports Medicine*, 13, 376-392.
- Mishra, M., & Ghosh, N. (2017). Development of performance in sprinting events in last 10 summer Olympics. *International journal of physical education, sports and health*, 4, 440-445.
- Morin, J., Edouard, P., Samozino, P. (2014). New Insights Into Sprint Biomechanics and Determinants of Elite 100m Performance. *New Studies in Athletics*, 3 / 4, 87-103.
- Niss, Mogens, & Blum, Warner. (2020). *The Learning and Teaching of Mathematical Modelling*. New York: Routledge.
- Prabowo, A., & Ngadiman. (2022). Mathematical Model for 100- and 200- meter Olympic Games Running Championship Time Records. *Mathematics and Statistics Journal* 10(6): 1141-1158.

- Supratman, A., Sobarna, A., & Rizal, R. M. (2024). Pengaruh Latihan Panjang Langkah dan Latihan Frekuensi Langkah terhadap Peningkatan Kecepatan Lari Cepat 60 Meter pada Siswa SMAN 1 Cisarua. *JIIP (Jurnal Ilmiah Ilmu Pendidikan)*, 7(6).
- 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.
- Valunpion, A., & Rangubhet, K. R. (2025). Performance and Equity in The 100-Meter Sprint: Contributions to Healthy and Inclusive Athletic Development. *Journal of Lifestyle& SDG'S Review*, 5, e04881. 01-19.
- Valamatos, M., Abrantes, J., Carnide, F., Valamatos, M., & Monteiro, C. (2022). Biomechanical Performance Factors in the Track and Field Sprint Start: A Systematic Review. *International Journal of Environmental Research and Public Health*, 19.
- Swinnen, W., Mylle I., Hoogkamer, W., Groote, F.D., & Vanwanseele W. (2021). Changing Stride Frequency Alters Average Joint Power and Power Distributions during Ground Contact and Leg Swing in Running. *Medicine & Science in Sports & Exercise*, 53(10), 2111-2118.