ComTech: Computer, Mathematics and Engineering Applications, 17(1), June 2026, 1−9 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v17i1.13224 E-ISSN: 2476-907X

Investigating Prospective Athletic Athletes: Classifiers, Benchmarking, and Post-Hoc XAI Analysis

Ibnu Febry Kurniawan[1]*; A'yunin Sofro[2]; Danang Ariyanto[3]; Junaidi Budi Prihanto[4]; Dimas Avian Maulana[5]

[1]Department of Data Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, 60231
[1]Innovation Center for Artificial Intelligence, Universitas Negeri Surabaya, Surabaya, 60213
[2,3,5]Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, 60231
[4]Department of Sport Education, Faculty of Sport Science and Health, Universitas Negeri Surabaya, Surabaya, 60213
[1]ibnufebry@unesa.ac.id; [2]ayuninsofro@unesa.ac.id; [3]danangariyanto@unesa.ac.id;
[4]junaidibudi@unesa.ac.id; [5]dimasmaulana@unesa.ac.id

Received: 19th March 2025/ **Revised:** 17th November 2025/ **Accepted:** 24th November 2025

How to Cite: Kurniawan, I. F., Sofro, A., Ariyanto, D., Prihanto, J. B., & Maulana, D. A. (2026). Investigating prospective athletic athletes: Classifiers, benchmarking, and Post-Hoc XAI analysis. *ComTech: Computer, Mathematics and Engineering Applications, 17*(1), 1−9. https://doi.org/10.21512/comtech.v17i1.13224

Abstract **-** Identifying highly potential athletes is a critical yet inherently challenging process that requires comprehensive analysis of diverse factors, including physiological attributes, demographic characteristics, and social influences. This multifaceted process requires meticulous evaluation of extensive datasets to ensure both accuracy and fairness in talent identification protocols. The complexity stems from the interconnected nature of the determinants of athletic performance, where physical capabilities intersect with psychological resilience, social support systems, and environmental factors. In recent years, machine learning (ML) algorithms gain prominence in decision-making processes, offering unprecedented opportunities to uncover subtle patterns and relationships within athlete data that might otherwise remain hidden. This study systematically benchmarks the performance of several state-of-the-art ML classifiers using a novel, self-collected dataset of athlete candidates. Furthermore, an explainable AI (XAI) technique, Shapley Additive Explanations (SHAP), is applied to interpret model decisions and provide meaningful insights into key predictive factors. Experimental results demonstrate that Gradient Boosting achieves superior predictive performance (F1) across the 10-fold sets, with a mean value of 0.46. SHAP analysis reveals the critical importance of anthropometric measurements and social group features in influencing prediction outcomes. These findings collectively underscore the substantial potential of ML to revolutionize talent identification in sports while emphasizing the importance of model interpretability in fostering trust and acceptance of AI-driven decision-making processes.

Keywords**:** machine learning, sports science, explainable AI, Post-Hoc analysis, benchmark

I. INTRODUCTION

The global sports industry's growing emphasis on evidence-based decision-making transforms talent identification and development. This transformation is particularly evident in the rapid adoption of advanced analytics and sophisticated data collection methods, which revolutionize how sporting organizations evaluate, develop, and value sports talent (Harde et al., 2025; Wrang et al., 2022; Zhang & Cao, 2025). Modern approaches incorporate a broader spectrum of characteristics that extend well beyond sports fields, such as demographic, social, and economic factors, to support athlete evaluation (Lu et al., 2023; Sofro et al., 2024). These factors contribute indirectly to health conditions such as hypertension and diabetes (Dey et al., 2022; Kabanda et al., 2022; Riddell et al., 2020; Schweiger et al., 2021), which influence an athlete's development trajectory, training adherence, and long-term performance sustainability (Alpsoy, 2020). The ability to systematically analyze these diverse factors alongside traditional athletic metrics represents a significant advancement in talent identification methodology.

Building on this multifaceted approach, machine learning (ML) applications in sports science demonstrate strong potential for processing complex, interconnected data and identifying subtle relationships among various athlete characteristics (Sharma et al., 2023). ML algorithms are successfully applied to performance analysis, injury prevention, and training

*Corresponding Author

optimization (Cesanelli et al., 2024; Wrang et al., 2022), and they show promise for understanding how socioeconomic and health factors interact with athletic development (Sofro et al., 2024). This capability to process and analyze multiple dimensions of athlete data simultaneously represents a significant advance beyond traditional statistical approaches.

Despite these technological advances, there remains a critical need for transparent and interpretable ML models in athlete selection processes. Current approaches often function as "black boxes" (Bodria et al., 2023; Hassija et al., 2024), which makes it challenging for sports practitioners to understand and trust the decision-making process. This limitation is particularly significant in high-stakes applications (Bodria et al., 2023), especially when evaluating prospective athletes with complex health considerations (Sharma et al., 2023), as organizations must clearly understand how various factors contribute to model predictions. The lack of interpretability poses a substantial barrier to the widespread adoption of ML tools in practical talent identification settings.

This research addresses these challenges by introducing a rigorous methodology that combines robust classifiers with an explainable artificial intelligence (XAI) technique for post-training (post-hoc) analysis of prospective athletes. No prior study evaluates ensemble classifiers while simultaneously providing interpretable insights and actionable recommendations for non-specialist stakeholders. The proposed approach is applied to a comprehensive dataset of prospective athletes that integrates traditional athletic metrics with socioeconomic and health-related factors. This pipeline leverages the strong predictive capabilities of ensemble methods while maintaining transparency through an XAI framework, enabling practitioners to understand how different factors influence athlete selection decisions.

II. METHODS

The dataset used in this study is collected through activities that record comprehensive measurements of 200 prospective athletes (Sofro et al., 2024). These measurements cover human physiology, socio-demographic, and health aspects, as described in Table 1. Descriptions of the coded features are provided in Table 2. The anthropometric section of the dataset includes physical measurements such as height, weight, waist circumference, and Body Mass Index (BMI). Meanwhile, the socio-demographic portion captures a combination of social

Table 1 Dataset Samples

#ID	TB	BB	IMT	Wt	A	G	EA	EI	JA	JI	SA	SI	F	Db	Hp	Atlet
R1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
R2	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0
R3	1	1	0	0	1	1	1	0	0	0	1	0	1	0	0	0

Table 2 Features and Their Brief Description

Code	Description	Categorization
TB	Height	[<170 cm, ≥ 170 cm]
BB	Weight	[< 60 kg, ≥ 60 kg]
IMT	BMI	[< 25, ≥ 25]
Waist circumference (Wt)	Waist length	[< 85 cm, ≥ 85 cm]
Age (A)	Age	[< 21, ≥ 21]
Gender (G)	Gender	[male, female]
Edu_Father (EA)	Father's Education Level	[school, college]
Edu_Mother (EI)	Mother's Education Level	
Job_Father (JA)	Father's Occupation Sector	[formal, informal]
Job_Mother (JI)	Mother's Occupation Sector	
Salary_1 (SA)	Father's salary level	[< 3, 3 – 6, ≥ 6] million IDR
Salary_2 (SI)	Mother's salary level	
Finance (F)	Family's overall financial level	[low, middle, high]
Diabetes (Db)	Diabetes status	[yes, no]
Hypertension (Hp)	Hypertension Status	
Atlet	Screening result (success/failure)	

and demographic attributes of respondents, including age, gender, parents' education, occupation, and salary. The final segment of the dataset records respondents' hypertension and diabetes test results.

In-person measurement and questionnaire sessions are conducted to record respondents' data. The data collection process ensures that all participant information is complete and consistently recorded. No additional preprocessing is performed aside from transforming the data into categorical variables. This categorization follows a standard procedure in which items are grouped based on shared characteristics or predefined criteria.

Several dataset samples are presented in Table 1, while the categorization criteria are outlined in Table 2. As shown in Table 2, standard athlete anthropometric measurements are used to represent respondents' physiological attributes, whereas general indicators are applied to capture socioeconomic characteristics. Of the 200 respondents, only 51 successfully become athletes, indicating a clear class imbalance within the dataset. This imbalance is quantified following Buda et al. (2018), with values of $\rho = 3.92$ and $\mu = 0.5$.

Next, this study employs a pipeline-based approach with particular attention given to data handling. The discriminative capability of learning algorithms, including machine learning (ML) models, is often sensitive to the data used during training and evaluation. Different subsets of data allocated for training, validation, and testing can introduce bias (Moreno-Torres et al., 2012). Consequently, variations in data splitting strategies at different stages of model development may lead to differing performance outcomes.

Considering these factors, the training mechanism in this study is designed primarily to capture and assess classifier performance on the dataset. The development steps presented in Figure 1 account for the intrinsic structure of the dataset and systematically evaluate classifier behavior. The pipeline begins with data collection and proceeds through model development and performance analysis. Since the analysis is conducted after model training, it is classified as a post-hoc analysis.

In many ML implementations, datasets are not strictly separated into fixed training and testing partitions. The underlying distribution of the target class presents an additional challenge, as the dataset is imbalanced. Training models on imbalanced data introduces bias toward the majority class, which can degrade both training and testing performance (Buda et al., 2018).

This study employs four ensemble learning algorithms due to their demonstrated robustness and strong predictive performance (Khan et al., 2024; Mienye & Sun, 2022). These algorithms construct multiple models during training and subsequently determine the optimal model based on residual or gradient optimization. Adaptive Boosting (AdaBoost, A), Gradient Boosting (G), and XGBoost (X) adopt sequential learning strategies, whereas Random Forest (R) applies a parallel learning strategy. A summary of these algorithms is provided in Table 3.

As described above, the training and testing process in this study is designed to capture classifiers' performance across different data splits and shuffling procedures. To support this objective, ten experimental sets with randomly shuffled training–testing partitions

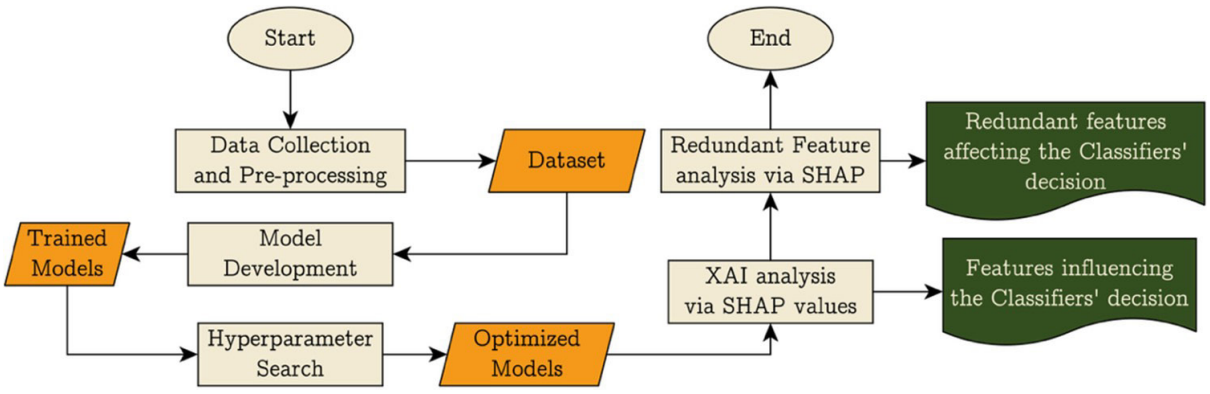

Figure 1 Methodology Used in this Study

Table 3 Core and Training Mechanism of Ensemble Learning Algorithms

Alg.	Type	Strat.	Learn Process	Error Correction
R	Bag	Par.	From a random bootstrap sample	-
A	Boost	Seq.	Focuses on misclassified samples	Reweight
G			Fit residual errors	Immediate prediction fit
X			Fit gradients loss	Gradient fit via Regularization

are generated. In each experimental set, 20% of the dataset is randomly assigned as test data. The remaining data are used for training through stratified K-Fold (SKFold) cross-validation, ensuring that class proportions are preserved across folds.

This fold arrangement preserves the target class distribution during cross-validation, ensuring that each training set contains a sufficient number of instances from each target class (Moreno-Torres et al., 2012; Szeghalmy & Fazekas, 2023). This design choice is particularly crucial for the present dataset due to its imbalanced class distribution. In addition, a nested configuration is employed as part of a robust model development strategy, in which a separate SKFold is prepared on the training set. This nested setup is used during cross-validation in conjunction with a hyperparameter search, as described in the subsequent section.

Another component of the training pipeline is model fine-tuning through hyperparameter search. All classifiers' hyperparameters, as shown in Table 4, are optimized using a grid search strategy with F1 score as the primary evaluation metric. This prioritization ensures the selection of models that balance precision and recall. Not all hyperparameters listed in Table 4 apply to every classifier due to inherent algorithmic differences. For example, the Random Forest classifier does not use a learning rate parameter, while the maximum depth configuration does not apply to the AdaBoost classifier. Hyperparameters marked with a tick (✓) indicate applicability to the corresponding classifiers. After the optimization process is completed, the most performant hyperparameter configuration is selected and used to train the full training set. The resulting trained classifier is then evaluated by generating predictions on the predefined test set.

Finally, a post-training (post-hoc) analysis is conducted on the trained models using the test set to assess performance during inference. SHapley Additive exPlanations (SHAP) (Lundberg et al., 2018) are applied to obtain insights into each feature's contribution to a classifier's prediction. This technique is grounded in the game-theoretic concept of Shapley values, which allocate a fair contribution to each feature based on its marginal impact. SHAP values quantify the difference between the expected model output and the actual prediction attributed to each feature. SHAP values for a given model *f(x)* with the model's baseline prediction ϕ_0 are presented in Equations (1) and (2) below.

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad (1)$$

$$\phi_0 = \mathbb{E}[f(X)] \quad (2)$$

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \left(\frac{|S|! \times (|N| - |S| - 1)!}{|N|!} \times [f(S \cup \{i\} - f(S))] \right) \quad (3)$$

Equation (3) computes the marginal effect of adding feature *i* to subset *S*. The weighted average of these marginal effects yields ϕi, the SHAP value for feature *i*, where *N* denotes the set of features in the dataset with cardinality *M*, and *S* represents a subset of *N* that excludes feature *i*. The resulting SHAP values are then used to interpret the classifier at both local and global levels.

Local explanations are obtained for individual predictions, enabling users to understand why the classifier produces a specific decision. This type of explanation is commonly visualized using force plots (Lundberg et al., 2018). In contrast, global explanations allow users to observe overall feature importance and identify recurring patterns in the interactions between the model and the dataset.

III. RESULTS AND DISCUSSIONS

The evaluation process begins by specifying metrics that measure classifier performance. Standard metrics are derived from fundamental evaluation outcomes, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values are then used to compute additional classification metrics, such as Precision (Prec), Specificity (TNR), Sensitivity (TPR), and F1.

The Precision metric in Equation (4) measures the proportion of relevant items among the retrieved items. Specificity, or the True Negative Rate shown in Equation (5), measures the proportion of correctly identified negative instances. Sensitivity, or the True Positive Rate presented in Equation (6), measures the proportion of relevant items correctly identified out of all relevant items. Finally, the F1 score, as shown in Equation (7), evaluates a classifier's predictive performance by combining Precision and Sensitivity into a single metric

$$\text{Precision (Prec)} = \text{TP}/(\text{TP} + \text{FP}) \quad (4)$$

$$\text{Specificity (TNR)} = \text{TN}/(\text{TN} + \text{FP}) \quad (5)$$

$$\text{Sensitivity (TPR)} = \text{TP}/(\text{TP} + \text{FN}) \quad (6)$$

$$F_1 = 2 \times \text{Prec} \times \text{TPR}/(\text{Prec} + \text{TPR}) \quad (7)$$

Table 4 Hyperparameters and Their Search Space during Fine-Tuning

Hyperparameters	Search Space	Classifier			
		A	G	R	X
n_estimators	[50, 100, 150, 200, 250]	ü	ü	ü	ü
learning_rate	[0.01, 0.05, 0.1 0.175, 0.2, 0.25]	ü	ü	û	ü
max_depth	[3, 4, 5, 6, 7, 8]	û	ü	ü	ü

Table 5 below summarizes each classifier's performance during testing. All values, reported as mean and standard deviation, are obtained from Stratified K-Fold cross-validation (K = 10). The highest metric achieved across the folds is indicated in brackets. The best-performing classifiers identified during the training phase are subsequently evaluated on the test data. The data partition varies across scenario sets, as described earlier. In each partition, the indices of the training and test data are randomized and recorded to ensure that each classifier is trained and evaluated on identical data splits, thereby supporting reproducibility. Several classifiers perform well in certain scenarios but exhibit weaker performance in others. This variability motivates the use of data index shuffling and systematic tuning procedures to obtain a more robust and reliable performance assessment.

According to Table 5, all classifiers demonstrate comparable performance across evaluation metrics and outperform the conventional classification approach reported by Sofro et al. (2024). The most performant classifiers developed in this study achieve improvements of approximately 2–3% in Accuracy, 3–30% in Precision, and up to 60% in Specificity. These results highlight the superior effectiveness of the ML algorithms, particularly when considering the extensive training scenarios applied in this study.

Among the evaluated classifiers, the smallest performance difference is observed in Accuracy, with a gap of 0.02 between the most and least performant models. On average, the Random Forest classifier attains an accuracy of 0.76, while AdaBoost achieves 0.74. The Random Forest classifier also exhibits a shorter interquartile range than AdaBoost, as presented in Figure 2, indicating that its accuracy values are more tightly clustered across experimental runs. This pattern suggests greater convergence in the Random Forest results, as the middle quartiles vary only slightly. However, this classifier also produces an outlier with an accuracy of approximately 0.68, indicating occasional performance degradation under specific data partitions.

Notable differences emerge in Specificity compared to Accuracy, as reflected by the varying lengths of the box plots. The gap between the highest and lowest mean Specificity values is approximately 0.06. Random Forest achieves the highest Specificity with an average of 0.97, whereas Gradient Boosting and XGBoost record lower mean values of 0.87. In contrast to the Accuracy metric, Random Forest does not exhibit any outlier results for Specificity. The high Specificity values indicate that the classifiers correctly identify the negative class with strong consistency. However, this strong performance in the True Negative Rate (TNR) is not matched by the corresponding True Positive Rate (TPR), or Sensitivity. All classifiers demonstrate difficulty in achieving satisfactory Sensitivity, with mean values ranging from 0.26

Table 5 Numerical Classifiers' Performance Over Test Sets.

Classifier		Metrics				
		Accuracy	Specificity	Sensitivity	Precision	F_1
AdaBoost	(A)	0.74 ± 0.03 (0.80)	0.89 ± 0.05 (0.97)	0.30 ± 0.13 (0.6)	0.49 ± 0.15 (0.75)	0.36 ± 0.12 (0.57)
GradientBoosting	(G)	0.75 ± 0.07 (0.82)	0.87 ± 0.08 (**1.0**)	**0.42** ± 0.16 (**0.7**)	**0.56** ± 0.23 (**1.0**)	**0.46** ± 0.15 (0.64)
Random Forest	(R)	**0.76** ± 0.04 (0.80)	**0.93** ± 0.04 (**1.0**)	0.26 ± 0.13 (0.6)	**0.56** ± 0.2 (**1.0**)	0.34 ± 0.13 (0.60)
XGBoost	(X)	0.75 ± 0.05 (**0.85**)	0.87 ± 0.05 (0.93)	0.40 ± 0.18 (0.7)	0.49 ± 0.13 (0.75)	0.43 ± 0.15 (**0.67**)

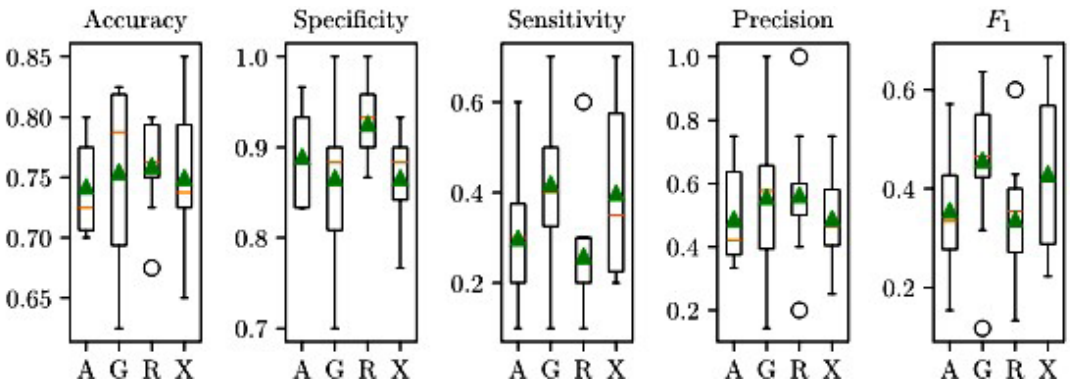

Figure 2 Boxplots for the Performance of Classifiers, i.e., AdaBoost (A), Gradient Boosting (G), Random Forest (R), and XGBoost (X).

to 0.42. These results suggest that class imbalance significantly influences classifier behavior, causing the models to favor the majority class, namely candidates who do not become athletes (Buda et al., 2018).

The F1 metric produced by all classifiers shows a noticeable performance gap. As shown in Table 5, the XGBoost classifier achieves the highest F1 score across the 10-fold cross-validation with a value of 0.67, followed closely by Gradient Boosting at 0.64. Despite XGBoost attaining the highest single score, Gradient Boosting is selected as the best-performing classifier due to its higher mean F1 score across all experimental runs. A higher mean F1 score indicates greater result stability, which leads to more reliable and dependable classifier performance.

The strong performance of Gradient Boosting reflects its inherent learning strategy, in which multiple weak learners are sequentially combined to form a strong learner through an iterative process. The algorithm employs an additive approximation approach that incrementally minimizes prediction error. Within this process, a weighting mechanism emphasizes harder-to-classify instances, enabling improved generalization performance (Mienye & Sun, 2022).

Moving on to model interpretability, this study employs local explanation analysis using SHapley Additive exPlanations (SHAP). As shown in Figure 3, a pair of SHAP explanations is presented for the trained Random Forest classifier. This illustration compares the classifier's decision for a single record in the test set and indicates whether the predicted target class is 0, meaning the prospect fails to become an athlete, or 1, meaning the prospect becomes an athlete. To achieve this level of detail, two force plots are provided, showing the contribution of all features to the prediction for target class 0 (Figure 3a) and target class 1 (Figure 3b). The ground-truth label for this instance is 0.

As shown in Figure 3a, the model output $f(x)$ equals 0.95, indicating a positive shift from the base value of approximately 0.72. In this visualization, the left-to-right arrow represents feature contributions that increase the prediction relative to the base value. In contrast, arrows pointing in the opposite direction in Figure 3b represent feature contributions that decrease the prediction from the base value. The base value, denoted as ($\mathbb{E}[f(x)]$), represents the classifier's expected output across the dataset for the given record. Since the ground-truth label for the testing instance in Figure 3a is 1, the force plot illustrates how individual features collectively push the prediction closer to that target outcome.

It is also worth noting in Figure 3a that all features are mirrored in Figure 3b. For example, the feature with the highest SHAP contribution, waist length (Wt), appears between the weight (BB) and Edu_Mother (EI) features in both visualizations. In addition, the length of each arrow reflects the magnitude of the corresponding feature's impact, where a stronger influence on the prediction results in a longer arrow. This visual consistency helps clarify how the same features contribute differently depending on the predicted class.

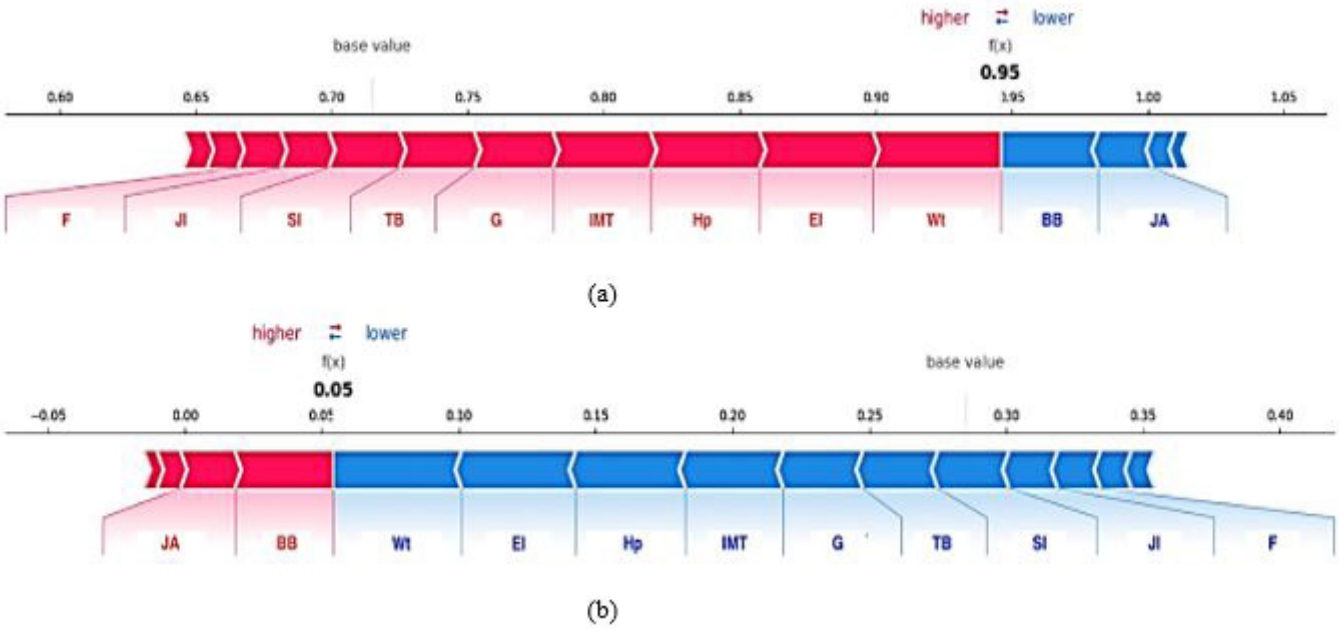

Figure 3 A Pair of Local Explanation Results Utilizing SHAP Force Plot on the Random Forest Classifier, Showing How the Trained Classifier Produces $f(x)$ for Target Class (a) 0 and (b) 1.

Next, this study conducts global explanation evaluation and feature importance assessment. As described in the model performance comparison section, multiple experiments are performed using shuffled training and testing indices. This index configuration ensures that all classifiers are evaluated under identical data conditions, enabling fair and reliable comparisons. The optimal hyperparameter configuration is then applied to each classifier prior to evaluation to ensure that performance and interpretability analyses reflect the best-performing model settings.

During the global explanation assessment, two models are selected based on overall performance rather than reliance on a single best-performing metric. XGBoost is selected because it achieves the highest F1 score in one of the ten test sets, indicating strong performance in balancing precision and recall. Meanwhile, Random Forest is chosen for its consistently strong results in Accuracy, Specificity, and Precision, despite exhibiting a lower mean F1 score. In addition, Random Forest is the only algorithm among the four that employs parallel model generation, which distinguishes its learning strategy from the other ensemble methods.

Figure 4 illustrates the global performance of both the XGBoost and Random Forest classifiers on the test data. Unlike the local explanation stage, this analysis focuses on capturing the overall behavior of each classifier as influenced by individual features across all test instances. The beeswarm plot presented in Figure 4 visualizes the distribution and magnitude of feature contributions for the entire dataset. Table 6 complements this visualization by summarizing the most influential features based on their mean absolute SHAP values (|SHAP value|), providing a concise overview of global feature importance.

In addition to the value assessments described above, this study uses a clustering mechanism applied during the testing phase in conjunction with SHAP values to identify redundant features. The clustering process is conducted to reveal inherent feature structure using a hierarchical method provided by the SHAP tool. This evaluation aims to determine which features in the dataset are relatively independent. It also investigates potential coupling or relationships between features that may lead to redundant contributions during model inference.

This process further reveals features that are closely related in terms of their contribution patterns. The threshold value, also referred to as the clustering cutoff, identifies pairs or groups of features with clustering distances no greater than the specified threshold. The clustering distance ranges from 0 to 1, where 0 indicates very close or duplicate features and 1 indicates complete independence. The

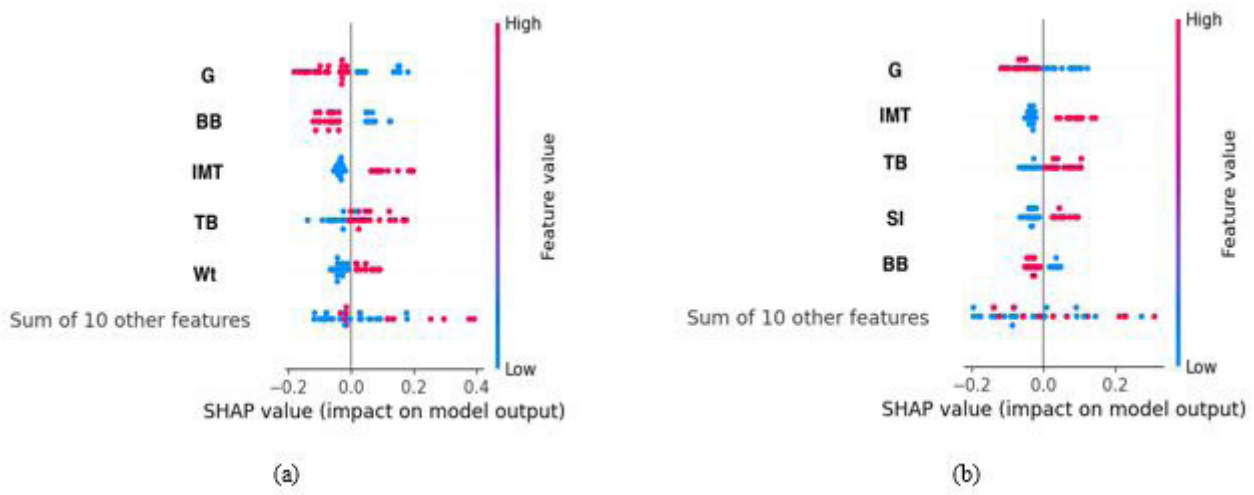

Figure 4 A Sample of Feature Importance Analysis on Trained Classifiers:
(a) XGBoost and (b) Random Forest

Table 6 Top 5 Features and Their Mean (|SHAP_value|).

Classifier	Feature				
	1	2	3	4	5
A	Wt (0.05)	TB (0.05)	BB (0.04)	EI (0.03)	Hp (0.02)
G	EA (0.12)	IMT	JA (0.07)	TB (0.06)	SA (0.06)
R	G (0.06)	TB (0.05)	IMT (0.05)	SI (0.04)	Wt (0.04)
X	G (0.03)	TB (0.08)	IMT (0.07)	BB (0.06)	Wt (0.05)

Table 7 Redundant Features with Clustering *Cut_Off* = 0.75

Classifiers	Feature	
	#1	#2
AdaBoost	N/A	N/A
Gradient Boosting	Finance	Gaji_2
Random Forest	Gaji_2	Finance
XGBoost	Gaji_2	Finance

clustering cutoff used in this study is set to 0.75, which represents a relatively conservative threshold close to independence. This threshold accommodates leaf clusters with a substantial margin and enables SHAP to uncover richer structural patterns within the classifiers.

According to Table 7, tree-based classifiers such as Gradient Boosting, Random Forest, and XGBoost indicate the presence of a feature pair labeled as redundant by SHAP. This labeling is obtained under the clustering cutoff of 0.75, which reinforces the conservative nature of the assessment. In contrast, AdaBoost identifies no redundant features within this cutoff range. This outcome suggests that, for AdaBoost, all key features contribute distinctly during the testing phase without exhibiting strong redundancy.

IV. CONCLUSIONS

This study examines the performance of several classifiers in predicting successful athletes based on social, demographic, and physical measurement records. Through nested stratified cross-validation, the study ensures a reliable and unbiased evaluation of model performance and finds that Gradient Boosting is the most consistent classifier across the test set, as indicated by the mean F1 score. This classifier also shows negligible differences across other key performance metrics.

Alongside this finding, this work utilizes an explainable AI tool. The use of SHAP for explainable AI provides valuable insights into model decisions, reveals critical factors that influence prediction outcomes, and enhances the interpretability of machine learning models. Experimental results highlight key features that influence predictions, particularly those dominated by the anthropometric group, such as Gender, Height, Weight, BMI, and waist length. Meanwhile, only a few classifiers consider demographic features, such as parents' occupation and salary, to be influential. Furthermore, only one classifier identifies the hypertension feature as a key determinant during the prediction phase. These contributions have practical implications for various stakeholders, including educational institutions and the sports industry, which seek data-driven approaches for talent identification.

Despite these contributions, this work is limited to data from the pre-selection process. Data collected during athlete admissions can reveal more valuable relationships between performance and outcomes. In addition, physiological measurements obtained during physical tests are missing, leaving the connections between athletes' physical condition and on-field performance underexplored.

To build on this work, it is recomended that future studies investigate various combinations of data types, including categorical and numerical features, along with additional feature groups to improve detection and provide a more comprehensive perspective. The incorporation of numerical data enables the observation and analysis of regression-based problems. In addition, integrating secondary data sources from similar sports serves as a substitute for missing data and increases the overall dataset size.

ACKNOWLEDGEMENT

This work was supported by Direktorat Riset Teknologi dan Pengabdian Masyarakat (DRTPM), Ministry of Education, Culture, Research and Technology of Indonesia, grant number 841/UN38/HK/2024.

AUTHOR CONTRIBUTIONS

Conceived and designed the analysis, I. F. K.; Collected the data, A. S., D. A. and J. B. P.; Contributed data or analysis tools, A. S., D. A. and D. A. M.; Performed the analysis, I. F. K.; Wrote the paper, I. F. K. And D. A. M.; Other contribution, A. S. and J. B. P.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [IFK], upon reasonable request. Explain the reason why the readers must request the data.

REFERENCES

Alpsoy, Ş. (2020). Exercise and hypertension. In J. Xiao (Ed.), *Physical exercise for human health*. Springer Nature Singapore. https://doi.org/10.1007/978-981-15-1792-1_10

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023). Benchmarking and survey of explanation methods for black box

models. *Data Mining and Knowledge Discovery*, *37*(5), 1719–1778. https://doi.org/10.1007/s10618-023-00933-9

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Cesanelli, L., Lagoute, T., Ylaite, B., Calleja-González, J., Fernández-Peña, E., Satkunskiene, D., Leite, N., & Venckunas, T. (2024). Uncovering success patterns in track cycling: Integrating performance data with coaches and athletes' perspectives. *Applied Sciences (Switzerland)*, *14*(7). https://doi.org/10.3390/app14073125

Dey, S., Mukherjee, A., Pati, M. K., Kar, A., Ramanaik, S., Pujar, A., Malve, V., Mohan, H. L., Jayanna, K., & N, S. (2022). Socio-demographic, behavioural and clinical factors influencing control of diabetes and hypertension in urban Mysore, South India: A mixed-method study conducted in 2018. *Archives of Public Health*, *80*(1), 234. https://doi.org/10.1186/s13690-022-00996-y

Harde, S., Bhawnani, V., & Savant, S. (2025). Comparative Analysis of data driven techniques to predict transfer prices of football players. *International Journal of Innovative Science and Research Technology*, 735–739. https://doi.org/10.38124/ijisrt/25mar351

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8

Kabanda, G. K., Nkodila, A. N., Masudi, G. M., Beya, F. E. B., Ngasa, N. N. K., Mety, R. M., Buila, N. B., Kayembe, J.-M. N., Longo, B. M., & M'Buyamba-Kabangu, J.-R. (2022). Impact of adapted physical activity on blood pressure and hypertension control in the militaries of Kinshasa garrison, Democratic Republic of Congo: A randomized controlled trial. *Annales Africaines de Medecine*, *15*(4), e4755–e4769. https://doi.org/10.4314/aamed.v15i4.2

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778. https://doi.org/10.1016/j.eswa.2023.122778

Lu, Y., Wiltshire, H. D., Baker, J. S., Wang, Q., & Ying, S. (2023). Associations between dairy consumption, physical activity, and blood pressure in Chinese young women. *Frontiers in Nutrition*, *10*, 1013503. https://doi.org/10.3389/fnut.2023.1013503

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.W., Newman, S.F., Kim, J., & Lee, S.I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Moreno-Torres, J. G., Saez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-Fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), 1304–1312. https://doi.org/10.1109/TNNLS.2012.2199516

Riddell, M. C., Scott, S. N., Fournier, P. A., Colberg, S. R., Gallen, I. W., Moser, O., Stettler, C., Yardley, J. E., Zaharieva, D. P., Adolfsson, P., & Bracken, R. M. (2020). The competitive athlete with type 1 diabetes. *Diabetologia*, *63*(8), 1475–1490. https://doi.org/10.1007/s00125-020-05183-8

Schweiger, V., Niederseer, D., Schmied, C., Attenhofer-Jost, C., & Caselli, S. (2021). Athletes and hypertension. *Current Cardiology Reports*, *23*(12), 176. https://doi.org/10.1007/s11886-021-01608-x

Sharma, S., Raval, M. S., Roy, M., Kaya, T., & Kapdi, R. (2023). Interpretable machine learning in athletics for injury risk prediction. In *Explainable AI in healthcare: Unboxing machine learning for biomedicine* (1st edn). Chapman and Hall/CRC. https://doi.org/10.1201/9781003333425

Sofro, A., Ariyanto, D., Budi Prihanto, J., A. Maulana, D., W. Romadhonia, R., & Maharani, A. (2024). Integration of bivariate logistic regression models and decision trees to explore the relationship between socio-demographic and anthropometric factors with the incidence of hypertension and diabetes in prospective athletes. *Sport Mont*, *22*(1), 71–78. https://doi.org/10.26773/smj.240210

Szeghalmy, S., & Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors*, *23*(4), 2333. https://doi.org/10.3390/s23042333

Wrang, C. M., Rossing, N. N., Agergaard, S., & Martin, L. J. (2022). The missing children: A systematic scoping review on talent identification and selection in football (soccer). *European Journal for Sport and Society*, *19*(2), 135–150. https://doi.org/10.1080/16138171.2021.1916224

Zhang, W., & Cao, D. (2025). Comparative analysis of hybrid and ensemble machine learning approaches in predicting football player transfer values. *Cognitive Computation*, *17*(2), 88. https://doi.org/10.1007/s12559-025-10443-z