ComTech: Computer, Mathematics and Engineering Applications, 17(1), June 2026, 1−9
DOI: 10.21512/comtech.v17i1.13224

P-ISSN: 2087-1244
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. (2025). 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 have gained 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. Further, 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 of 0.46. SHAP analysis reveals the critical importance of anthropometric measurements and social group features in influencing prediction outcomes. These findings collectively underscore ML's substantial potential to revolutionize talent identification in sports while emphasizing the paramount 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 now incorporate a broader spectrum of characteristics that extend well beyond sports fields, such as demographic, social, and economic factors, to assist athlete evaluation (Lu et al., 2023; Sofro et al., 2024). These factors contribute to an indirect influence on health conditions such as hypertension and diabetes (Dey et al., 2022; Kabanda et al., 2022; Riddell et al., 2020; Schweiger et al., 2021), which impact 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 remarkable 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,

*Corresponding Author

and training optimization (Cesanelli et al., 2024; Wrang et al., 2022), and they show strong 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), making it challenging for sports practitioners to understand and trust the decision-making process. This limitation is particularly significant in highly critical applications (Bodria et al., 2023), especially when evaluating prospective athletes with complex health considerations (Sharma et al., 2023), as organizations need to clearly understand how various factors contribute to the model's 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 XAI technique for post-training (post-hoc) analyses of prospective athletes. To the best of our knowledge, no prior study conducts evaluations over ensemble classifiers to provide interpretation and recommendations to non-specialists. Our approach is applied to a comprehensive dataset of prospective athletes that uniquely combines traditional athletic metrics with broader socioeconomic and health metrics. The proposed pipeline leverages the superior predictive capabilities of ensemble methods while maintaining transparency through an XAI tool, thereby 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 readings cover human physiology, socio-demographic, and health aspects, as described in Table 1. Descriptions for 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 and demographic attributes

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)	

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. Each team member surveys the participants, ensuring that the process yields complete data. No further preprocessing is performed aside from transforming the data into categorical variables. This categorization is a standard process in which items are grouped based on common criteria.

Several dataset samples are provided in Table 1, and the categorization criteria are outlined in Table 2. As shown in Table 2, standard athlete anthropometric measurements are used for respondents' physiological measurements, while general metrics are applied for socioeconomic measurements. Overall, 51 of the 200 respondents successfully become athletes, indicating a class imbalance within the dataset. This imbalance is quantified at a level (Buda et al., 2018) of $\rho = 3.92$ and $\mu = 0.5$.

Next, this study employs a pipeline-based approach, with special attention given to the data. The discriminative capability of learning algorithms, including ML, is often sensitive to the data that are fed and used during their assessment. Different subsets of data used for training, validation, and testing contribute to bias (Moreno-Torres et al., 2012). Hence, applying different data splits at each stage of model development produces varying performance outcomes.

Considering the factors mentioned above, the training mechanism conducted in this study is primarily designed to gather classifiers' performance on the dataset. The development steps (see Figure 1) account for the dataset's intrinsic structure and capture the classifiers' performance. The pipeline starts from data collection and proceeds through model development and analysis. Since the analysis is conducted after training, it is considered a post-hoc analysis.

In most ML implementations, the datasets used do not have a strict split between training and testing. The inherent distribution of the target class poses an additional challenge, as the dataset is imbalanced. Training on an imbalanced dataset introduces bias toward the majority class, which results in lower overall training and testing performance (Buda et al., 2018).

This study employs four ensemble learning algorithms, primarily due to their robust performance (Khan et al., 2024; Mienye & Sun, 2022). These algorithms generate multiple models during training and then select the optimal one based on the model's residual or gradient. Adaptive Boosting (AdaBoost, A), Gradient Boosting (G), and XGBoost (X) algorithms use a sequential learning strategy, while Random Forest (R) adopts a parallel strategy. A summary of these algorithms is presented in Table 3.

As described above, the training and testing process employed in this work is designed to capture classifiers' performance across different data splits and shuffling procedures. To achieve this goal, ten sets of experiments with shuffled training–testing partitions are prepared. In each experiment set, 20% of the dataset is assigned as test data using a random process. The remaining portion of the dataset is used for training with stratified K-fold (SKFold) cross-validation.

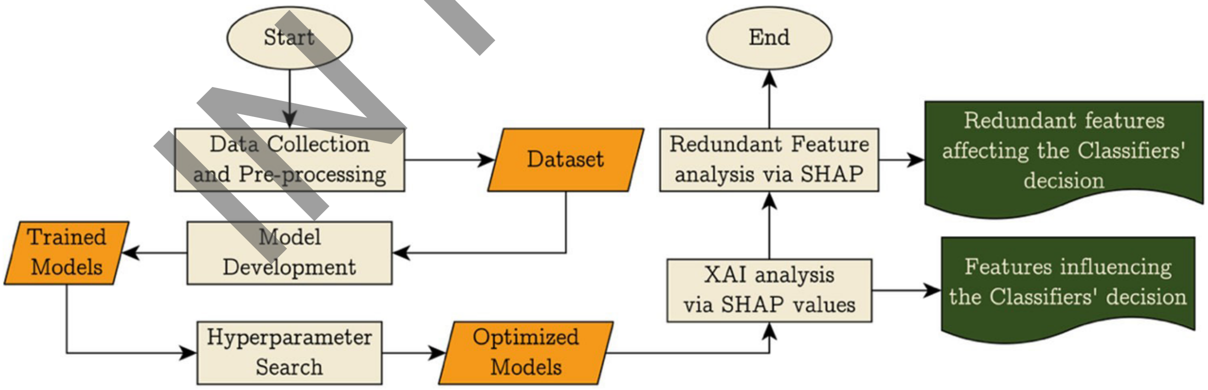

Start
Data Collection and Pre-processing
Dataset
Model Development
Trained Models
Hyperparameter Search
Optimized Models
XAI analysis via SHAP values
Redundant Feature analysis via SHAP
End
Redundant features affecting the Classifiers' decision
Features influencing the Classifiers' decision

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

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 choice is particularly crucial for the present dataset due to its imbalanced nature. Next, this study employs a nested configuration as an additional setup for robust model development, in which a separate SKFold is prepared on the training set. This additional setup is used in cross-validation alongside a hyperparameter search, and the process is detailed in the subsequent section.

Another component in the training pipeline is model fine-tuning, or hyperparameter search. All classifiers' hyperparameters (see Table 4) are optimized using a grid search with F1 as the primary metric. This prioritization aims to select the best-performing classifier that balances precision and recall. Not all hyperparameters defined in Table 4 apply to every classifier due to operational differences. For instance, the Random Forest classifier does not accommodate the learning rate hyperparameter, while the max depth configuration does not apply to the AdaBoost classifier. Hyperparameters with tick (✓) apply to the corresponding classifiers. Upon completing the process, the most performant hyperparameter combination is selected and is used to train the entire training set. The trained classifier is then evaluated by obtaining predictions on the predetermined test set.

Finally, a post-training (post-hoc) analysis is conducted on the trained models using the test set to assess their performance during inference. SHapley Additive exPlanations (SHAP) (S. Lundberg & Lee, 2017) are used to obtain insights into each feature's contribution to a classifier's prediction. This tool is grounded in a game-theoretic concept, namely Shapley values, which allocate a fair contribution to each feature. SHAP values measure 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 predictio ϕ_0 are shown in Equations (1) and (2) below.

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

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

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

Eq. 3 computes the marginal effect of adding i to S. Then, the weighted average of these marginal effects yields ϕi, the SHAP value for feature i, *where* N is the set of features in the dataset, with cardinality M, and S is a subset of N excluding feature i. The SHAP values can then be applied to interpret the classifier locally and globally. The local explanation is obtained for individual predictions, enabling users to understand why the classifier produces a decision. This approach is commonly represented in force plots (S. M. Lundberg et al., 2018). Meanwhile, the global explanation helps users observe overall feature importance and patterns occurring in model-dataset interaction.

III. RESULTS AND DISCUSSIONS

The evaluation process begins with specifying metrics to measure how the classifiers perform. Standard metrics stemming from fundamental evaluations, such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values are then used in further classification metrics, such as Precision (Prec), Specificity (TNR), Sensitivity (TPR), and F1. The Prec (Eq. 4) metric measures the proportion of relevant items among the retrieved items. Specificity, or TN Rate (Eq. 5), measures the ratio of relevant items that are not returned. Sensitivity or TP Rate (Eq. 6) measures the ratio of relevant items returned from all relevant ones. Finally, F1 (Eq. 7) measures the predictive performance of a classifier, accounting for Prec and TPR.

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

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

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

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

Table 5 summarizes each classifier's performance during testing. All values, formatted in mean and standard deviation, are obtained from Stratified-K Fold (K=10). The highest metric achieved across K is in brackets. The best-performing classifiers from the training phase are then evaluated on the test data. The data partition, as described, varies across scenario sets. In each partition, the indices of training and test data are randomized and recorded (1) to ensure

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

each classifier trains and tests the same set of data and (2) for reproducibility. Several classifiers may perform best in several scenarios but struggle in others. Such a condition motivates the use of the data indices shuffle and tuning process.

According to Table 5, all classifiers show similar performance across metrics and outperform the conventional classification approach presented by Sofro et al. (2024). Most performant classifiers developed in this study yield 2-3% higher Accuracy, 3-30% higher Precision, and 60% higher Specificity. These results highlight the superior performance of the ML algorithms, considering the extensive training scenarios employed in this study.

Among the classifiers, the smallest difference is in accuracy, with a gap of 0.02 between the most- and least-performant classifiers. On average, the Random Forest classifier produces 0.76, while AdaBoost can produce 0.74. The former classifier provides a shorter interquartile range than the latter (see Figure 2). This indicates that Random Forest models' accuracy tends to converge, i.e., the middle quartiles of experimental results differ slightly. However, this classifier produces an outlier, showing roughly at 0.68.

Next, there are notable differences in Specificity compared to accuracy, as the boxes in the plots differ in length. The highest and lowest mean Specificity differ by approximately 0.06. Random Forest achieves the best Specificity, averaging 0.97, while Gradient Boosting and XGBoost yield 0.87. In contrast to the Accuracy metric, Random Forest does not have any outlier results. This high Specificity suggests that all classifiers correctly classify the negative class. However, these good True Negative Rate (TNR) results are not followed by their counterpart, True Positive Rate (TPR) or Sensitivity. All classifiers struggle to achieve satisfactory results, with the lowest mean of 0.26 and the highest of 0.42. These exhibits suggest that class imbalance affects classifiers' performance, leading them to lean towards one class (Buda et al., 2018), i.e., prospects failing to become athletes.

The F_1 metric produced by all classifiers shows a noticeable gap. As seen in Table 5, the XGBoost classifier achieves the highest score across the 10-fold with 0.67, comparable to Gradient Boosting with 0.64. Despite the XGBoost's performance, it is preferable to choose Gradient Boosting as the most-performing classifier due to its higher mean F_1 score. The higher the mean F_1 score, the more likely the classifier is to produce consistent results, leading to dependable performance.

Further, the Gradient Boosting achievement reflects its inherent strategy, in which weak learners are sequentially combined to form a strong learner through an iterative process. The algorithm utilizes an additive approximation. Therein, a weighting mechanism is employed to obtain more accurate generalization (Mienye & Sun, 2022).

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**)

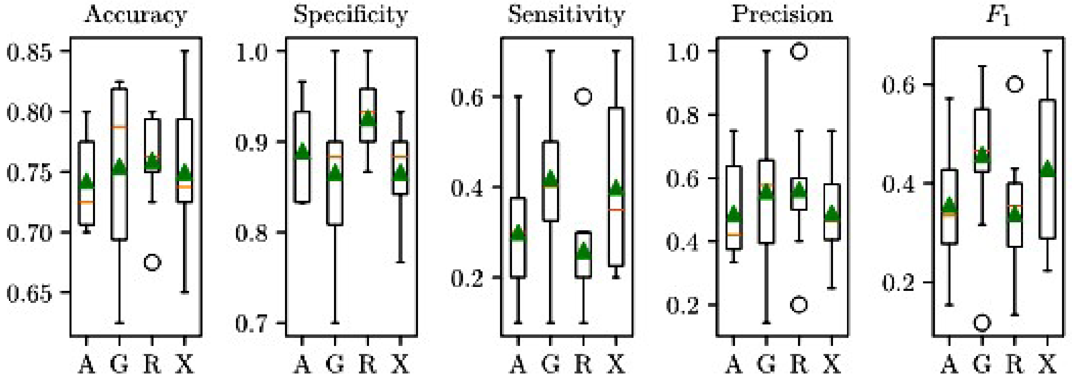


Figure 2 Boxplots for the Performance of Classifiers, i.e., AdaBoost (A), Gradient Boosting (G), Random Forest (R), and XGBoost (X).

Moving on to the models' interpretability, this study employs local explanation activity using SHAP. As seen in Figure 3, a pair of SHAP values on the trained Random Forest classifier. This illustration compares the classifier's decision for a record in the test set and indicates whether the predicted target class is 0 (prospect failed to become an athlete) or 1 (otherwise). To obtain this level of detail, a pair of force plots showing the contribution of all features to the decision to target class 0 (see Figure 3a) and class 1 (see Figure 3b) is presented. The ground truth for this instance is 0.

As seen in Figure 3a, $f(x)$ = 0.95, indicating a push from the base value ≈ 0.72. Therein, the arrow from left to right indicates increments to the base value. Likewise, the opposite direction in Figure 3b shows decrements made to the base value. The base value ($\mathbb{E}[f(x)]$) represents the classifier's average (expected) output across the dataset for the corresponding record. Since the ground-truth label for the testing data in Figure 3a is 1, the force plot illustrates the features' contributions, leading to a prediction close to the label (1).

It is also worth noticing in Figure 3a that all features are mirrored to those in Figure 3b. For instance, the feature contributing the highest SHAP value, i.e., Waist length (Wt), is placed between Weight (BB) and Edu_Mother (EI) feature. Also, the length of each arrow reflects the impact made by the corresponding feature, where the bigger the impact, the longer the arrow's length.

Next, this study conducts Global Explanation evaluation and feature importance assessment. As described in the Model Performance Comparison part, there are several experiments, each with shuffled training and testing indices. The index setup ensures that all classifiers receive the same data configuration, ensuring fair comparisons. Then, the optimal hyperparameter configuration is applied to classifiers before conducting evaluation.

During the global explanation assessment, two models are selected based on their performance, not solely on the best-performing aspect. XGBoost is chosen because it achieves the highest F_1 score across 1 of 10 test sets. Meanwhile, Random Forest is chosen for its performance on Accuracy, Specificity, and Precision, but it falls short on the mean F_1 score. Besides, Random Forest is the only algorithm out of the four with parallel model generation.

Figure 4 illustrates the global performance of both the XGBoost and Random Forest classifiers on the test data. In contrast to the local explanation step, this step attempts to capture the overall classifier behavior contributed by each feature. Hence, the beeswarm plot shown in Figure 4 shows all features and their contribution in each test data entry. Table 6 complements this information by summarizing key features based on the mean (|SHAP value|).

In addition to the value assessments above, one can use a clustering mechanism considered during the testing phase, along with SHAP values, to identify redundant features. The clustering herein is conducted

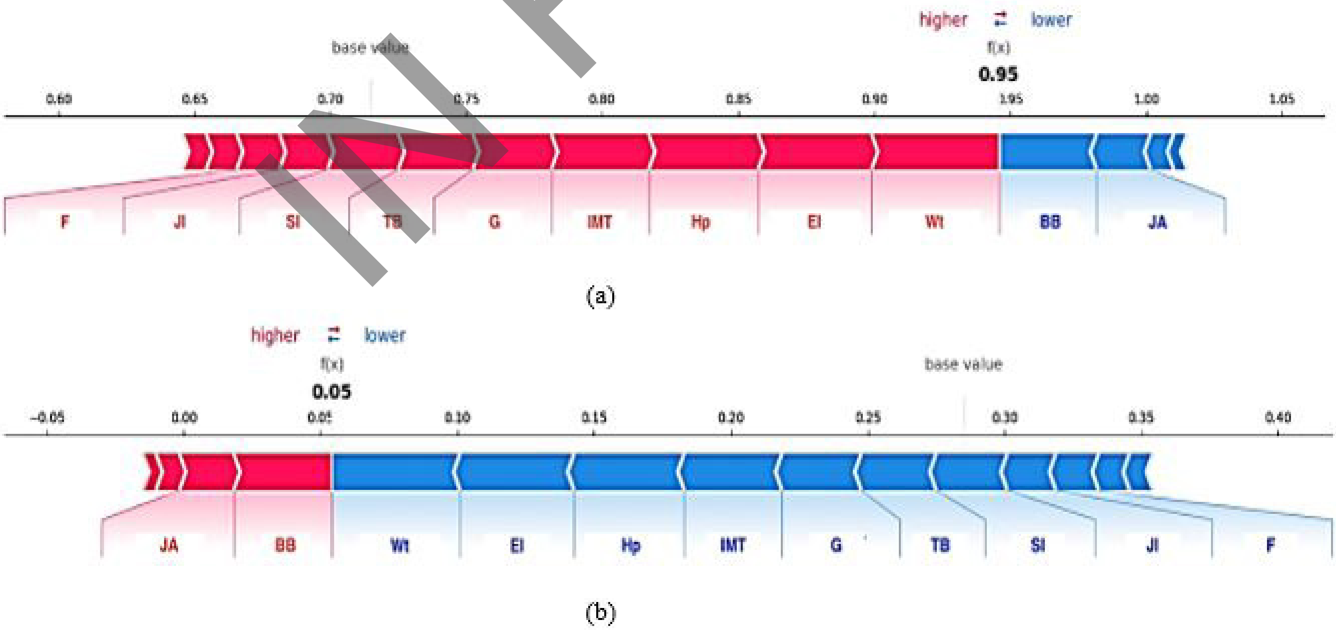
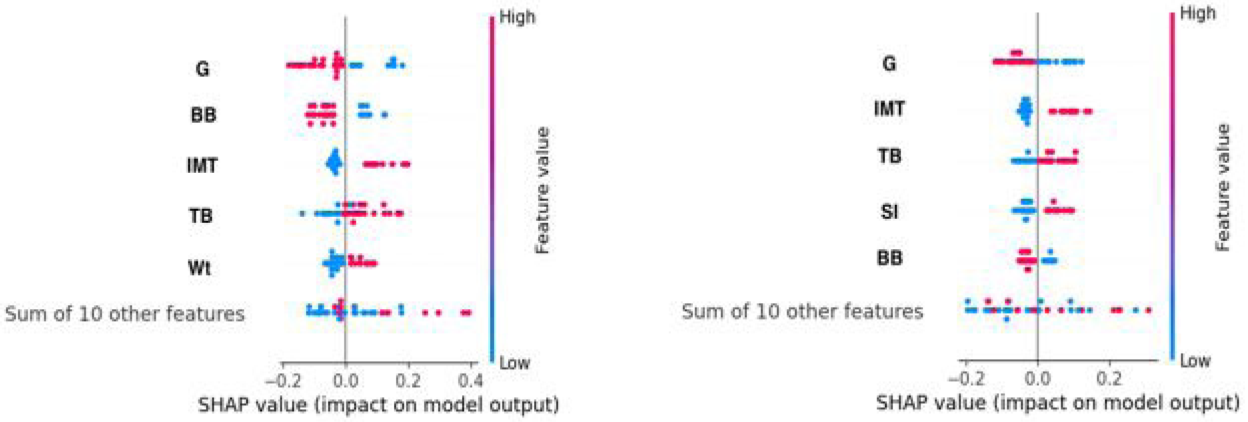


Figure 3 A Pair of Local Explanation Results Utilizing SHAP Force Plot on the Random Forest Classifier, Showing How the Trained Classifier Produces $f(x)$ for Target Class (a) 0 and (b) 1.



(a) (b)

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

to reveal inherent structure via a hierarchical method provided by the SHAP tool. This evaluation aims to identify which features in the dataset are relatively independent. Also, this process investigates any coupling/relation between features, leading to redundant contributions.

Alternatively, such a process reveals which features are closely related. The threshold value called clustering cut-off, shows pairs or groups of features with clustering distance up to the threshold value. The clustering distance ranges from 0 to 1, where 0 means very close or duplicate, and 1 indicates independent. The clustering cut-off value herein is 0.75 because no pairs were found during our evaluations. Such a value is relatively conservative, since it is close to 1. The value accommodates leaf clusters with a fairly large margin. This setup also allows SHAP to reveal more structure in classifiers.

According to Table 7, tree-based classifiers, such as Gradient Boosting, Random Forest, and XGBoost, indicate that a pair of features is labeled as redundant by SHAP. However, this label is obtained by setting the clustering cut-off to 0.75, which is conservative. On the other hand, AdaBoost returns no redundant features that fall within this threshold. This implies that the classifier considers all key features during testing.

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, 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* (Vol. 1228, pp. 153–167). 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