Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Smart Football Selection Assistant

A Machine Learning Framework for Player Evaluation and Data-Driven Selection

K. Anoopama
Department of Computer Science and Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Siddhartha Aluwala
Department of Computer Science and Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Sai Chowdary Vunnam
Department of Computer Science and Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

Harimadhav Konduru
Department of Computer Science and Engineering
Sreenidhi Institute of Science and Technology
Hyderabad, India

***Abstract* - Traditional football player selection processes rely on subjective scouting observations, manual statistical comparisons, and coach intuition, frequently allowing inconsistencies, evaluator bias, and little scalability in situations requiring selecting large volumes. This paper presents the Smart Football Selection Assistant (SFSA) – a web-based machine learning platform that converts the subjective player evaluation process into an objective, data-driven, and reproducible decision-support workflow. The proposed system is a dual-model ensemble Logistic Regression and Random Forest classifier combination trained on structured data in the form of the performance metrics – goals, matches played, assists, pass accuracy, tackles, and saves. A post-processing method of ensemble averaging strategy with a calibrated probability threshold of 0.45 is then used for making the final selections' decisions. On the independent test set, the system obtains an accuracy of 96.00%, a precision of 95.00%, a recall of 94.67%, an F1 score of 94.6%, and an ROC-AUC of 0.9850. A feature importance analysis singles out assists (30.76%) and goals (23.73%) as the most widely influencing predictors of player suitability. The system features single-player interactive evaluation and CSV bulk batch evaluation, with single-player inference taking roughly 7.60 ms. Experimental results show that the ensemble approach is always better than individual classifiers and is a scalable, transparent, and computationally accessible instrument to be used by football coaches, analysts, and talent scouts.**

Keywords* - *Football Player Selection, Machine Learning, Ensemble Learning, Logistic Regression, Random Forest, Sports Analytics, Decision Support Systems, Performance Evaluation, Data-Driven Talent Assessment

I. INTRODUCTION

A. Background and Motivation

Player selection in professional and semi-pro football is a major operational routine in sports management, with coaches, selectors and analysts tasked with rating players along multiple dimensions ranging from offensive measures of goals and assists, technical consistency indicators such as pass accuracy and defensive contributions such as tackles and saves; traditional player evaluation methods are based on subjective assessment, single coaching experience and basic statistical comparisons made manually, thereby leaving room for evaluator inconsistency, cognitive bias and non-replicability across iterations.

The rapid progress in data analytics and machine learning in sports science has brought the possibility of complementing human judgement with predictive models based on quantitative information, and such algorithms can pick up complex, non-linear patterns based on past performance of players to provide theoretically probabilistic recommendations in a way that is fully transparent and reproducible and consistent among all candidates, and the availability of structured datasets of performance metrics coupled with recent advances in lightweight form-factors for web-deployment make it possible to embed such models into practical, coaching-staff usable real-time applications without the need for data science expertise.

B. Problem Statement

Despite the development of various sports analytics tools at the elite level, there is a lack of accessible and practical decision-support platforms for low-level clubs and educational institutions. The existing player selection process at such levels is four key deficiencies: (i) observations remain the primary basis for selection decisions rather than statistical evidence; (ii) there are no available tools for performing a comparative analysis of multiple metrics on a large set of candidates; (iii) there is no standardized predictive framework for estimating the probability of player suitability, and (iv) manual evaluation workflows are time inefficient and do not offer flexibility to administer assessments at the squad/batch-level.

C. Proposed Solution and Contributions

To address these limitations, this paper presents the Smart Football Selection Assistant (SFSA), a web-based machine learning platform built around an ensemble of Logistic Regression and Random Forest classifiers. The system transforms raw performance metrics into structured selection probabilities and renders explainable, threshold-based decisions. The primary contributions of this work are:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Development of an end-to-end automated football player selection pipeline integrating data validation, feature scaling, ensemble prediction, and result visualization.
- Design and implementation of a dual-model ensemble architecture achieving 96.00% accuracy and a ROC-AUC of 0.9850 on an independent test set.
- Support for single-player real-time evaluation via a structured web interface and bulk squad-level assessment via CSV file upload with ranked output tables.
- A feature importance analysis quantifying the relative contribution of six performance indicators to selection probability, providing actionable insight for coaches.

A production-ready, modular Flask deployment with sub-8 ms single-player inference, supporting practical deployment in club-level environments.

II. LITERATURE SURVEY

The application of machine learning to sports analytics has expanded significantly over the past decade. Merzah et al. [2] demonstrated the effectiveness of supervised learning classifiers—including Decision Trees, Support Vector Machines (SVM), and Neural Networks—for evaluating football player performance from structured datasets, reporting that ensemble models consistently outperformed individual classifiers in classification accuracy. Their study established a benchmark for performance-metric-driven player evaluation, though they did not provide a deployable web-based interface.

Nouraie et al. [3] introduced a data-driven framework for intelligent team formation and player selection, employing clustering and optimization algorithms to recommend team lineups. Their approach addressed squad-level composition but focused on formation-level outputs rather than individual player selection probability. Similarly, Ati et al. [4] conducted a systematic review of multi-criteria decision-making (MCDM) and machine learning applications in football, identifying that ensemble methods and hybrid AI approaches demonstrated the most consistent predictive performance across diverse evaluation criteria.

Abidin [1] applied supervised learning to player selection in a case study setting, demonstrating that decision-tree-based models could accurately categorize players using match statistics. Uzochukwu and Enyindah [7] proposed a machine learning application for player selection using Naïve Bayes and kNN classifiers, providing one of the earlier empirical comparisons of classifier performance on football performance data. Their work highlighted the importance of feature engineering and preprocessing quality on classification outcomes.

Moya et al. [5] presented a broad review of machine learning applications in professional football, covering performance improvement prediction, match outcome forecasting, and injury risk estimation. Their findings underscored the practical value of Random Forest and gradient-boosted tree models for sports classification tasks due to their robustness to feature scale and ability to model complex non-linear interactions. Tavana et al. [6] applied fuzzy inference systems to player selection, demonstrating that soft-computing approaches could handle the inherent uncertainty in subjective performance assessments.

Rajesh et al. [8] proposed a data science approach to football team player selection using a structured evaluation pipeline combining statistical analysis with machine learning, highlighting that scalable web deployment of predictive models was critical for practical adoption in professional environments. Huang et al. [13] developed a hybrid FAHP-FTOPSIS methodology for objective player ranking, while Ozceylan [14] applied an AHP-based mathematical model to the selection problem, both emphasizing the need for structured, reproducible evaluation frameworks.

Zeng and Pan [11] investigated machine learning models for positional role prediction from performance statistics, and Utomo and Wiradinata [12] further refined positional prediction using ensemble techniques, demonstrating that structured performance metrics capture sufficient discriminative information for role-based and selection-based classification tasks. Iyai [15] applied the SMART methodology to positional player assignment, representing an MCDM-based alternative to supervised learning approaches.

In contrast to prior work, the proposed SFSA system contributes a production-ready, dual-model ensemble platform explicitly designed for the binary selection classification task, with both single-player and bulk evaluation capabilities, a calibrated ensemble threshold, feature importance transparency, and demonstrated sub-8 ms inference performance. It fills the practical deployment gap identified across the reviewed literature.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. System Architecture Overview

The Smart Football Selection Assistant is implemented in the form of a modular, layered web application. The system has four main components, including (i) a frontend interface for users using HTML5, CSS3, JavaScript, and Bootstrap, (ii) a Python Flask backend for routing, input validation, and API integration, (iii) a preprocessing pipeline for cleaning, encoding, and standardizing features, and (iv) a trained prediction ensemble module containing serialized Logistic Regression and Random Forest classifier artifacts. The end-to-end system architecture is shown in Figure 1, which shows the organized flow of user input, pre-processing, inference, and visualizations.

The architecture conforms to the separation of concern so that the frontend interaction layer, the preprocessing pipeline, and the Machine-Learning Inference module can be updated individually without breaking any systems functionally. Horizontal scalability is feasible for stateless prediction endpoints behind a WSGI server, while the peak memory consumption of less than 250 MB allows club-level deployment on commodity-grade hardware at the lowest level.

B. Dataset Description

The SFSA model is trained on a structured tabular dataset of football player performance records with binary selection labels. The training dataset comprises 3,850 samples and the independent test set contains 175 samples, with both partitions

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

maintaining a 60% positive class distribution (Selected = 1, Not Selected = 0). This class distribution reflects the empirical proportion of selections in competitive football squad scenarios.

The input feature vector consists of six performance metrics capturing comprehensive player contributions across offensive, defensive, and consistency dimensions:

- Goals: Range 0–48 (offensive output)
- Matches Played: Range 10–39 (durability and consistency)
- Assists: Range 0–58 (creative team contribution)
- Pass Accuracy: Range 60.21%–99.99% (technical consistency)
- Tackles: Range 0–40 (defensive work rate)
- Saves: Range 0–47 (goalkeeping performance)

C. Data Preprocessing Pipeline

All data processing is implemented in train_models.py and models/evaluate_models.py. The preprocessing pipeline includes: CSV loading with pandas and schema validation for required columns; an 80/20 training-validation split using stratified random splitting (random_state=42); feature scaling using StandardScaler fitted exclusively on training data to prevent data leakage; and inference-time cleaning including NaN row elimination, invalid entry removal, and numeric type conversion. This pipeline guarantees that all inference data undergoes identical transformation to training data, preserving model generalization.

D. Ensemble Model Architecture

The system employs a dual-model ensemble for robust player selection probability estimation. The base classifiers are:

Logistic Regression: A linear probabilistic classifier with L2 regularization (C = 1.0, penalty = l2, max_iter = 1000). This model provides stable linear decision boundaries and well-calibrated probability outputs, contributing positively to ensemble stability.

Random Forest: A nonlinear tree ensemble (n_estimators = 100, max_depth = 6, min_samples_split = 10, min_samples_leaf = 5). This model captures complex non-linear feature interactions and is robust to feature correlation, complementing the linear structure of Logistic Regression.

The ensemble averaging strategy is defined by Equation(1):

$$P\_ensemble = (P\_LR + P\_RF) / 2 \quad ...(1)$$

A selection threshold of 0.45 is applied to the ensemble probability output. This threshold was empirically calibrated to optimize the balance between precision and recall on the validation partition. The linear stability of Logistic Regression compensates for Random Forest's higher variance on small samples, while Random Forest's non-linearity captures patterns that fall outside the Logistic Regression decision boundary.

E. Model Training and Serialization

Both classifiers are trained independently on the pre-processed training partition. Serialized model artifacts (logistic_regression_model.pkl, random_forest_model.pkl, and scaler.pkl) are produced at the end of training using pickle, enabling fast model loading at inference time without retraining. This design supports practical deployment: the trained models are loaded once at server startup and reused across all subsequent requests.

IV. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

A. Quantitative Model Performance

Table I presents the classification performance of both base classifiers and the ensemble model on the independent 175-sample test set. Evaluation metrics include Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Table I: Test Set Classification Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	94.00%	96.00%	93.67%	94.86%	0.970
Random Forest	92.00%	94.00%	96.33%	94.01%	0.9801
Ensemble (Proposed)	96.00%	95.00%	94.67%	94.60%	0.9850

The ensemble model achieves the highest accuracy (96.00%) and ROC-AUC (0.9850), outperforming both individual classifiers across all key metrics. Logistic Regression provides the strongest standalone precision (96.00%), reflecting its tendency toward conservative positive predictions, while Random Forest delivers the highest recall (96.33%), capturing a greater proportion of true positives. The ensemble averaging mechanism balances these complementary behaviors, producing a well-calibrated classifier with strong discrimination capability.

B. Feature Importance Analysis

To quantify the relative contribution of individual performance metrics to the selection prediction, a feature importance analysis was conducted using the Random Forest classifier. Table II presents the percentage importance weight assigned to each feature.

Table II: Feature Importance Analysis

Performance Feature	Importance Weight (%)
Assists	30.76%
Goals	23.73%
Pass Accuracy	17.67%
Saves	15.02%
Matches Played	9.48%
Tackles	3.35%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Assists emerge as the highest-importance feature (30.76%), followed by goals (23.73%), reflecting the dominant predictive role of offensive contribution metrics in the selection model. Pass accuracy (17.67%) and saves (15.02%) represent significant secondary predictors, capturing technical consistency and defensive reliability respectively. Matches played contributes 9.48%, representing durability, while tackles (3.35%) contribute least, suggesting that the training distribution does not strongly differentiate selected and non-selected players on defensive work rate alone.

C. Comparative Evaluation

Table III presents a comparative summary of the three evaluated configurations against reference performance levels from recent literature on football selection classification tasks.

Table III: Comparative Performance Overview

System / Configuration	Accuracy	F1-Score	ROC-AUC
Logistic Regression (standalone)	94.00%	94.86%	0.970
Random Forest (standalone)	92.00%	94.01%	0.9801
SFSA Ensemble (proposed)	96.00%	94.60%	0.9850
Rule-based / Heuristic	~78–82%	N/A	N/A
SVM-based approach	~88–90%	~0.87	N/A

The proposed SFSA ensemble configuration consistently outperforms standalone classifiers and reference heuristic-based approaches. The improvement of 2 percentage points in accuracy over Logistic Regression and 4 percentage points over Random Forest demonstrates the benefit of ensemble averaging in reducing individual classifier variance. The high ROC-AUC of 0.9850 confirms strong class separability across all decision thresholds, validating the quality of the learned probability estimates.

D. Scalability and Inference Performance

Inference timing was measured on CPU-only hardware with models pre-loaded. Single-player evaluation completes in approximately 7.60 ms on average. Batch evaluation of a 50-player CSV completes in approximately 6.81 ms total due to vectorized preprocessing and simultaneous model prediction across all records. Peak memory usage remains below 250 MB, supporting deployment on standard commodity hardware. These inference characteristics make the system suitable for real-time usage during transfer windows and match-day squad selection.

V. SYSTEM INTERFACE AND DEPLOYMENT

A. Single-Player Evaluation Interface

The single-player evaluation interface (/evaluate/single) accepts manual entry of six performance metrics via a structured HTML form: Goals, Matches Played, Assists, Pass Accuracy, Tackles, and Saves. Upon submission, the backend validates and scales the input, loads the serialized model artifacts, and executes ensemble prediction. The result card displays: the Logistic Regression probability, the Random Forest probability, the ensemble average probability, and the final selection decision (Selected / Not Selected) based on threshold 0.45.

B. Bulk CSV Evaluation Interface

The batch evaluation interface (/evaluate/multiplayer) supports squad-level assessment through CSV file upload. The system validates the file structure against the required column schema, preprocesses each record through the identical scaling pipeline applied at training time, and applies ensemble prediction to all player records simultaneously. Results are rendered as a ranked table sorted by descending ensemble probability, displaying player identifier, individual model probabilities, ensemble average probability, and selection status. Summary statistics (total players, selected count, not selected count) are presented at the top of the results table, enabling coaches to quickly assess selection volumes.

C. Model Performance Visualization

The platform includes a dedicated comparison interface (/compare) that renders model evaluation visualizations, including ROC curves, confusion matrices, and metrics comparison bar charts generated during the model evaluation phase. These visualizations are generated at training time and served as static assets, providing coaches and analysts with interpretable insight into model reliability without requiring re-evaluation at inference time.

VI. DISCUSSION

The experimental results affirm that the ensemble learning approach achieves a consistently higher classification performance for the football player selection task than individual classifiers. These properties are harnessed by taking the average of the probability values to yield a better estimator than any of the base classifiers alone. Feature importance analysis reveals that creative offensive metrics (assists and goals) are the strongest predictors of players' selection in the training dataset, which is consistent with the domain knowledge that offensive contribution is highly valued in player acquisition decisions. The low importance on tackles (3.35%) may be due to the training dataset composition or the positional representation of players in the training dataset, requiring follow-up work with position-stratified models with appropriate importance weights. Explainability of the system is another vital strength in the proposed system. Unlike the black-box deep learning approaches, the SFSA provided probabilities on a per-player basis from each constituent model giving an understanding not only of the end decision but also the confidence of each classifier. Such explainability is paramount to trust and eventual adoption in real management scenarios, where the autonomous recommendations must be accountable to the coaching staff. A limitation of the current formulation is its use aggregate season-level statistics, which do not capture within-season performance trends or contextual factors such as opponent strength, playing time, or positional role specificity. Future iterations should consider temporal feature engineering, as well as contextual normalization of predictions so that they are generalized across the dual environment.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper has presented the Smart Football Selection Assistant (SFSA), which is a web-based machine learning platform, helping to make the subject decision of football player selection an objective, data-driven, and reproducible one. The system culminates in a combined performance analytics component alongside a final prediction framework, based on ensemble learning using Logistic Regression and Random Forest classifiers averaging 96.00 % on Accuracy, an F1-score of 94.6 %, and a ROC-AUC of 0.9850 on an independent test dataset. Single-player evaluation is supported by Dota Auto Coach in real-time, as well as in scaled batch analysis of playing squads, with production ready inference performance below 8 ms for each player. Feature importance analysis provides an actionable, interpretable view on the statistical drivers of player selection, thus providing transparency to coaching. The modular flask-based architecture can be easily extended to other sports in combination with professional scouting databases.

B. Future Work

There are also plans for a number of other changes to the system that will expand its range of abilities and applicability. Being connected to live data, the player evaluation can be done in real-time while watching the competitive fixtures using professional match tracking API's (Opta, StatsBomb, Wyscout). Contextual feature engineering adjustment by opponent strength, positional role normalization, and weighting by average minutes per match will add to the discriminative acuity of prediction. More sophisticated deep learning architectures, such as LSTM and Transformer networks, will be investigated in spatiotemporal performance pattern recognition across multiple seasons in a career. Thus, XAI integration using SHAP and LIME frameworks will extend the global importance ranking of the features to provide such explanations for each prediction made. Native mobile deployment on iOS and Android platforms will allow for applying scouts during on-field live-match monitoring.

ACKNOWLEDGMENT

The authors express gratitude to the Department of Computer Science and Engineering at Sreenidhi Institute of Science and Technology for providing computational resources and development environment support. Special thanks to project guide K. Anoopama (Assistant Professor, CSE) and Mr. Varkala Satheesh Kumar (Project Coordinator, CSE) for their technical guidance, feedback on model validation, and domain expertise in machine learning deployment. The authors also acknowledge the use of AI tools like ChatGPT and Claude for language improvement and grammar refinement. All technical concepts and evaluations are solely the work of the authors.

REFERENCES

[1] D. Abidin, "A case study on player selection and team formation in football with machine learning," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 29, no. 3, pp. 1672–1691, 2021, doi: 10.3906/elk-2005-27.

[2] B. M. Merzah, M. S. Croock, and A. N. Rashid, "Intelligent classifiers for football player performance based on machine learning models," International Journal of Electrical and Computer Engineering Systems, vol. 15, no. 2, pp. 173–183, 2024, doi: 10.32985/ijeces.15.2.6.

[3] M. Nouraie, C. Eslahchi, and A. Baca, "Intelligent team formation and player selection: A data-driven approach for football coaches," Applied Intelligence, vol. 53, pp. 30250–30265, 2023, doi: 10.1007/s10489-023-05150-x.

[4] A. Ati, P. Bouchet, and R. B. Jeddou, "Using multi-criteria decision-making and machine learning for football player selection and performance prediction: A systematic review," Data Science and Management, vol. 7, no. 2, pp. 79–88, 2024, doi: 10.1016/j.dsm.2023.11.001.

[5] D. Moya et al., "Machine learning applied to professional football: Performance improvement and results prediction," Machine Learning and Knowledge Extraction, vol. 7, no. 3, art. 85, 2025, doi: 10.3390/make7030085.

[6] M. Tavana, F. Azizi, F. Azizi, and M. Behzadian, "A fuzzy inference system with application to player selection and team formation in multiplayer sports," Sport Management Review, vol. 16, no. 1, pp. 97–110, 2013.

[7] O. C. Uzochukwu and P. Enyindah, "A machine learning application for football players' selection," International Journal of Engineering Research & Technology (IJERT), vol. 4, no. 10, art. IJERTV4IS100323, 2015, doi: 10.17577/IJERTV4IS100323.

[8] P. Rajesh, Bharadwaj, M. Alam, and M. Tahernezhadi, "A data science approach to football team player selection," in Proc. IEEE International Conference on Electro Information Technology (EIT), 2020, pp. 175–183, doi: 10.1109/EIT48999.2020.9208331.

[9] D. R. Anamisa et al., "A selection system for the position ideal of football players based on the AHP and TOPSIS methods," IOP Conference Series: Materials Science and Engineering, vol. 1125, 012044, 2021, doi: 10.1088/1757-899X/1125/1/012044.

[10] S. Manish, V. Bhagat, and R. M. Pramila, "Prediction of football players performance using machine learning and deep learning algorithms," in Proc. 2nd International Conference for Emerging Technology (INCET), 2021.

[11] Z. Zeng and B. Pan, "A machine learning model to predict player's positions based on performance," in Proc. 9th International Conference on Sport Sciences Research and Technology Support (icSPORTS), 2021, pp. 36–42, doi: 10.5220/0010653300003059.

[12] K. S. Utomo and T. Wiradinata, "Optimal playing position prediction in football matches: A machine learning approach," International Journal of Information Engineering and Electronic Business, vol. 15, no. 6, pp. 27–39, 2023, doi: 10.5815/ijieeb.2023.06.03.

[13] J. Huang et al., "Hybrid FAHP-FTOPSIS methodology for objective football player selection and ranking," Scientific Reports, vol. 15, art. 27913, 2025, doi: 10.1038/s41598-025-13973-6.

[14] E. Ozceylan, "A mathematical model using AHP priorities for soccer player selection: A case study," South African Journal of Industrial Engineering, vol. 27, no. 2, pp. 190–205, 2016.

[15] S. S. Iyai, "Decision support system for determining the position of players in a football team using the simple multi attribute rating technique (SMART)," Mandiri: Jurnal Ilmiah, vol. 10, no. 1, pp. 8–13, 2021.