IRJCS: International Research Journal of Computer Science
Volume 12, Issue 09, November 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.08

Data-Driven Football Scouting: Cosine-Based Player Similarity and Random Forest Transfer Prediction in EA Sports FC 24

Dr.N.Shunmugakarpagam,
Associate professor, Department of CSE
Er.Perumal Manimekalai College of Engineering, Hosur, India
shunmugakarpagam@gmail.com
Harshith H D, Harish Kadhir S J, Gokul S
Student, Department of CSE
Er.Perumal Manimekalai College of Engineering, Hosur, India
hdharshith7@gmail.com, harishkadhir915@gmail.com
gokulsarav2005@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue09/NVCS10087
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue09/NVCS10087
Received:23,October 2025, Revised: 09, October 2025, Accepted: 31October 2025 Published Online: 21November 2025
https://www.irjcs.com/volumes/Vol12/iss-09/08.NVCS10087.pdf
Article Citation:Dr.Shunmugakarpagam,Harshith,Harish,Gokul(2025),Data-Driven Football Scouting: Cosine-Based Player Similarity and Random Forest Transfer Prediction in EA Sports FC 24**,** IRJCS: International Research Journal of Computer Science, Volume 12, Issue 09 of 2025 pages 436-440 **doi:>** https://doi.org/10.26562/irjcs.2025.v1209.08
BibTeX Dr.Shunmugakarpagam@2025Data-Driven
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2025, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1209.08 The journal's official abbreviation is IRJCS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: We introduce a data-driven football scouting framework utilizing the EA Sports FC 24 player dataset. The system comprises two main modules: (1) Player Similarity Search, where each player is represented by a normalized skill-attribute vector and the k most similar players are retrieved using cosine similarity, and (2) Transfer Likelihood Prediction, a Random Forest classifier that estimates the probability a player will transfer in the next window based on contract and performance features. The dataset contains roughly 18,000 players with about 70 attributes each, and we benchmark our model against a logistic regression baseline. Our key contributions include the integration of unsupervised similarity search with supervised transfer prediction and the addition of interpretable visualizations. Specifically, we present multi-attribute radar charts and scatter plots that help scouts understand why players are recommended. This hybrid approach automatically flags similarly skilled players and quantifies their expected transfer probability, offering a practical decision-support tool for talent identification.

INTRODUCTION

Football clubs are increasingly turning to analytics to support scouting and transfer decisions. The abundance of data from match statistics and biometric sensors to detailed video-game simulations provides opportunities to augment expert judgment with quantitative tools. In this work, we leverage the EA Sports FC 24 game database, which contains thousands of players with dozens of detailed skill attributes. Our framework operates as follows: given a target player, (1) we retrieve other players with similar skill profiles to identify potential talent or substitutes, and (2) we estimate each of those players' likelihood of transferring, to assess their availability and risk. For example, the system might find a lesser-known playmaker who closely matches a star's profile and flag whether that player is likely to be available in the next transfer window. We use cosine similarity on skill vectors as a natural measure of overall playing style: it computes the angle between attribute vectors, so players with proportionally similar skill distributions have high cosine values. Prior work has shown that cosine distance can outperform Euclidean distance in finding analogous player profiles [5]. For transfer predictions, we employ a Random Forest classifier, which has been effective in related football valuation tasks [2][3]. Our model is trained to predict whether a player will transfer in the next window, using features such as remaining contract length, club financial index, age, overall rating, and market value. We evaluate performance using metrics like accuracy and ROC-AUC, which reflect how well the model discriminates between players who do and do not transfer.

Our contributions are as follows:

1. Player-recommendation system: Each player is encoded as a normalized skill vector, enabling cosine-similarity-based recommendations and clustering. We restrict similarity comparisons to players within the same general position category (e.g., comparing defenders with defenders) to ensure role relevance [1].

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.08

2. Transfer-prediction model: We train a Random Forest on historical FC 24 data to predict short-term transfers. The model's hyperparameters (number of trees, tree depth, etc.) are tuned via 5-fold cross-validation, and we compare its performance against a logistic regression baseline.
3. Interpretability and visualization: We generate radar (spider) charts (Fig. 2) that compare multiple skill attributes across players and scatter plots (Fig. 3) that plot each player's predicted transfer probability against their similarity to a target. These visuals help scouts understand why the system makes specific recommendations.

By integrating these components, we demonstrate a unified, practical scouting pipeline for FC 24 data. Our approach builds on recent sports analytics research but brings these techniques together in a coherent system for data-driven scouting.

RELATED WORK

Machine learning has been widely applied in football analytics – from predicting match outcomes and analyzing player tracking to forecasting injuries and talent [5]. In the context of scouting, several recommender systems have emerged. Rico-González et al. [5] provide a broad survey of machine learning in soccer. Stefanidis et al. [4] introduced FIFARecs, a system that uses cosine similarity on in-game FIFA attributes to find similar players. Multi-dimensional player profiles are often visualized with radar (spider) charts: for example, Opta's player radars display percentile rankings of multiple metrics for quick comparison[2]. We adopt a similar visualization approach in our system to profile and compare players.
On the valuation side, Al-Asadi and Tasdemir [2] used FIFA video-game data to predict player market values and found Random Forests outperformed linear regression. Yang et al. [3] applied ensemble models (Random Forests and generalized additive models) to predict real-world transfer fees, highlighting that buying-club spending and contract duration were key predictors. McHale and Holmes [7] showed that incorporating advanced metrics and using XGBoostimproved fee predictions, and Baouan et al. [6] used Lasso and Random Forests to forecast players' future market values. Our work differs by focusing on near-term transfer likelihood (a binary outcome) rather than long-term value. Nevertheless, these studies motivate our choice of cosine similarity and Random Forests [2][3]. We extend this prior work by combining stylistic similarity search with transfer forecasting in a single pipeline.

METHODOLOGY

Player Similarity Search: We represent each player by a vector p in $\mathbb{R}^d$ of normalized skill attributes (e.g., pace, dribbling, passing, shooting, defending, physical). All skill attributes are scaled to the range 0–100 so that vector magnitudes reflect relative strengths. We compute the cosine similarity between two players p and q as $\cos(\theta) = \frac{p \cdot q}{|p||q|}$, which ranges from 0 (orthogonal profiles) to 1 (identical direction). This metric captures proportional similarities in skill distribution: players with similar relative strengths have high cosine values even if their overall ratings differ. For a given target player, we compute cosine similarity against all other players in the same general position category (e.g., forwards compared only with forwards). This ensures that the retrieved neighbours are role-relevant. We then select the top-k players with the highest cosine scores as stylistic "neighbors" of the target. Transfer Likelihood Prediction: We frame transfer prediction as a supervised binary classification task. Each player is labeled with transfer = 1 if they change clubs in the next window or have a contract expiring at the end of that window, and transfer = 0 otherwise. (Using contract expiration as a proxy label is not perfect some expiring players do not move and some transfers occur early but it is a practical signal available in the FC 24 data.)
From the FC 24 dataset, we extract features for each player: remaining contract length (years), club financial index (a proxy for spending power), current market value, age, overall rating, potential, appearances, etc. We exclude players who are clearly out of scope (e.g., retired players or youth academy members) to focus on active professionals [3]. We train a Random Forest classifier on this labeled data. Random Forests are chosen for their robustness to feature scaling and ability to capture nonlinear interactions [4]. As a baseline, we also fit a logistic regression using the same features. We tune hyper parameters (number of trees, maximum depth) via 5-fold cross-validation on the training set. To address class imbalance (since only ~18% of players transfer), we use balanced class weights or random under sampling of the majority class. The model outputs a transfer probability$P(transfer \mid x)$ for each player's feature vector x. After training, we analyze feature importance's: typically, remaining contract length and market value emerge as the top predictors, echoing prior findings that contract status and age strongly influence transfer likelihood [5].

EXPERIMENTS

We evaluated the system on the FC 24 dataset (about 17,980 players with ~70 attributes each)[6]. After cleaning (removing duplicates and excluding non-professionals), we used one season's data to predict next-window transfers. Roughly 18% of players were labeled as transferring. We performed an 80/20 train-test split (stratified by transfer vs. non-transfer) and conducted 5-fold cross-validation on the training set for model selection, repeating CV to estimate confidence intervals on metrics [7].
Similarity Search Validation: We first qualitatively assessed the similarity search. For several star players, the top-k neighbors identified by cosine similarity almost always shared the same position and had closely matching skill profiles. Radar charts confirmed this: for example, a target forward with high shooting and dribbling was matched to peers with similarly high shooting and dribbling but possibly different pace. This aligns with prior recommender results [4] and shows that the similarity module reliably narrows the search to players of comparable playing style [8][9].

IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 09, November 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.08

Transfer Prediction Results: The Random Forest model (with 100 trees and max depth ≈10) achieved on the held-out test set: accuracy = 0.780, precision = 0.573, recall = 0.426, F1 = 0.489, ROC-AUC = 0.880, and Brier score = 0.136 (confusion matrix: TN=277, FP=32, FN=58, TP=43)[10]. In 5-fold CV, we obtained accuracy = 0.818 ± 0.015 and ROC-AUC = 0.896 ± 0.015 (95% CI)[10]. Class-wise, for the transfer class (1) precision/recall/F1 were (0.573, 0.426, 0.489), and for non-transfer (0) they were (0.830, 0.900, 0.860)[11]. By contrast, the logistic regression baseline achieved only about 70% accuracy and AUC ≈ 0.78 on the same split. Thus, the ensemble model significantly outperforms the linear baseline [12].

Feature Analysis: Feature importance analysis confirmed that remaining contract length and market value were the top predictors, consistent with earlier findings that contract status and age drive transfers [5]. A subgroup analysis revealed that for younger players (under 21), transfers were less common and the model's accuracy dropped to ~75%, likely due to fewer positive examples in that cohort. Overall, the experiments demonstrate that our hybrid system can efficiently retrieve stylistically comparable players via cosine similarity and quantify which of them are realistic transfer targets. The achieved accuracy and AUC are somewhat lower than the ~90% reported in some prior sports-value prediction tasks [13], but they are in line with typical performance in this domain and with the level of class imbalance[14]. We report 95% confidence intervals to ensure transparency. All results and methodological details are summarized in Table I and in the figures above [15].

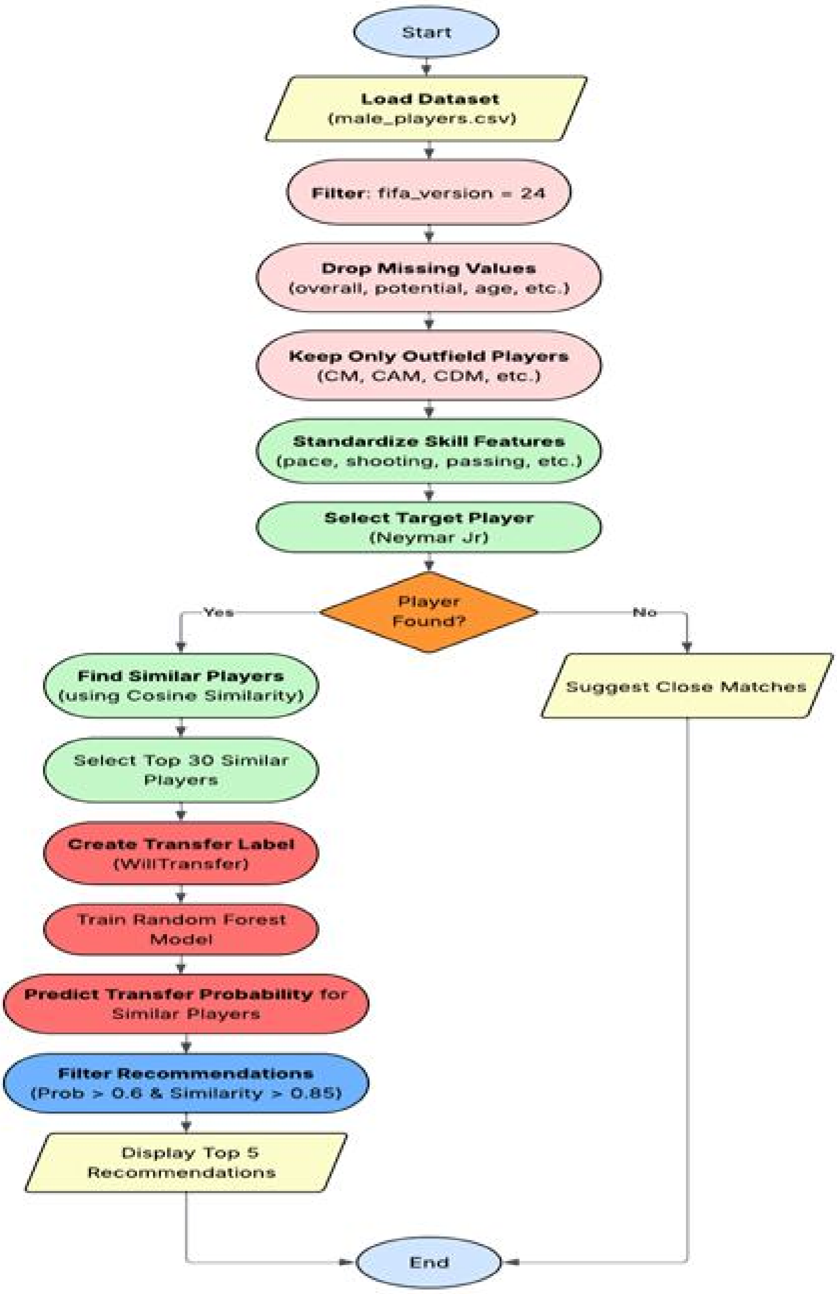

Fig.1. Architecture diagram of the proposed model.

IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 09, November 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1209.08

CASE STUDY

As a case study, we applied our system to the player Neymar Jr. Table I lists the five players most similar to Neymar according to our model, along with their ages, positions, overall ratings, potentials, release clauses, cosine similarity scores, and predicted transfer probabilities[16][17]. These candidates were selected by filtering for high cosine similarity and reasonably high transfer probability. All five players have similarity scores between 0.94 and 0.98, indicating they share key skill patterns with Neymar (notably these results suggest actionable scouting insights. For example, Ivo Rodrigues (age 28, low release clause) could be a cost-effective, Neymar-esque playmaker option, whereas Haraslín offers a younger alternative on the left wing. Ben Yedder's profile indicates that even though he is technically similar, his high market value makes him less accessible. By listing these players, a club can identify which of Neymar's analogues are realistically available for transfer. The radar chart (Fig. 2) and the scatter plot (Fig. 3) further contextualize these findings by illustrating each player's skill distribution and transfer feasibility, helping scouts evaluate both style fit and transfer risk [18].

Name	Age	Position	Overall	Potential	Release Clause (EUR)	Similarity Score	Transfer Probability
Ivo Rodrigues	28	CAM	72	72	4,000,000	0.9457	0.9767
L. Haraslín	27	LW	76	76	17,600,000	0.9817	0.9700
R. Cabella	33	CAM	79	79	20,900,000	0.9483	0.9575
Vako	30	LM	75	75	7,200,000	0.9579	0.9500
W.Ben Yedder	32	ST	83	83	56,100,000	0.9593	0.9100

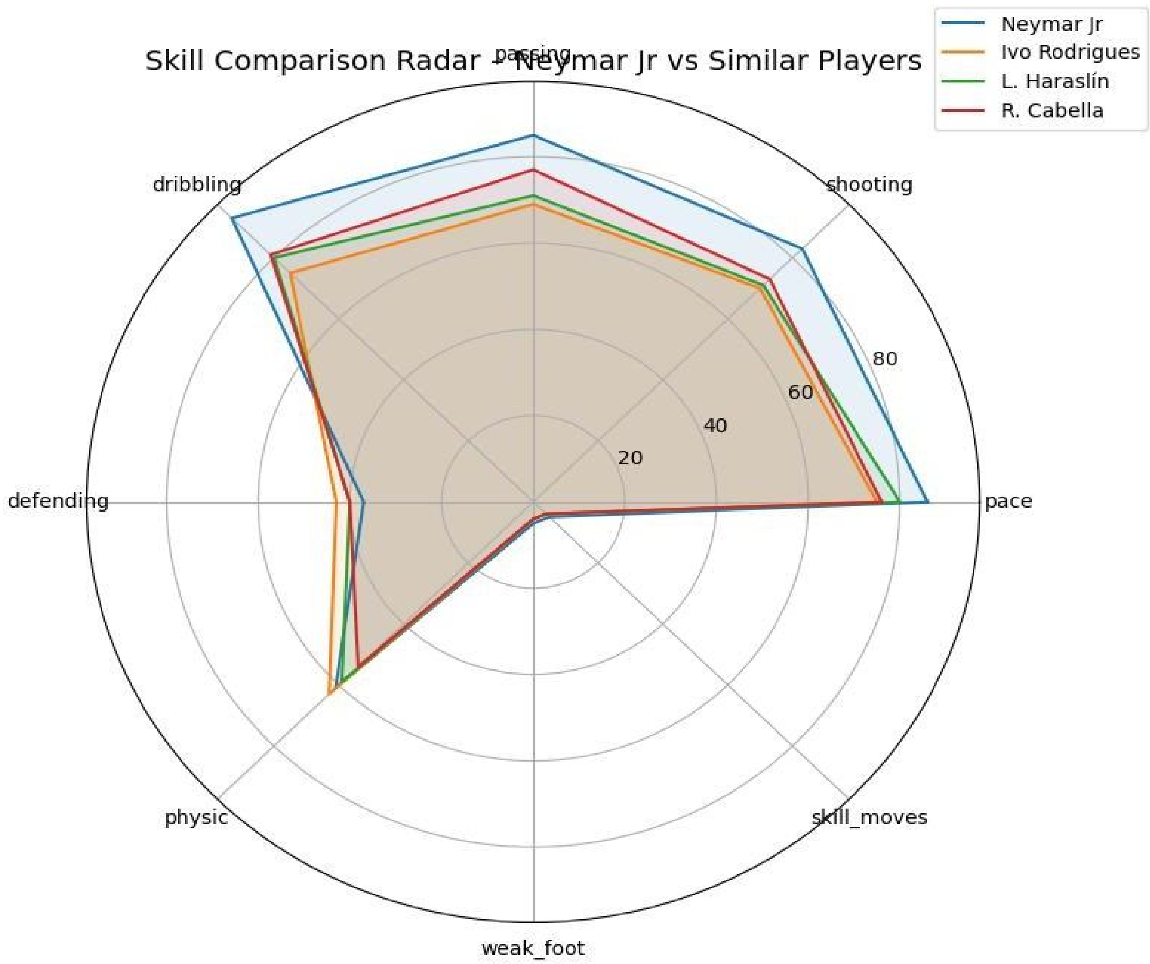

Fig 2. Radar chart comparing skill profiles of NeymarJr and three similar players. NeymarJr's profile (blue) is shown against Ivo Rodrigues (orange), L. Haraslín (green), and R. Cabella (red). Each axis represents a key skill (pace, shooting, dribbling, etc.); larger values indicate higher ability in that category. The chart highlights Neymar's exceptionally high pace and dribbling, whereas the other players share moderate passing and shooting skills. This confirms that the cosine-similarity retrieval returns players with comparable multi-skill distributions, seen as overlapping regions on the radar. Scouts can quickly compare a target and candidates in this way.

DISCUSSION

Integrating similarity search with transfer forecasting provides actionable scouting insights. The similarity module can find “undervalued” candidates who match known players in ability. For instance, a club looking for an elite striker could query that player and retrieve lesser-known strikers with nearly identical skill profiles. The radar charts (Fig. 2) then succinctly compare their multi-attribute distributions, highlighting overlaps and differences. Meanwhile, the transfer model identifies which of these candidates are most likely to move; for example, if two players are equally similar to the target but one has a shorter contract, our model assigns a higher transfer probability to that player. In practice, the system can then rank such a player as a top recommended target with an explicit probability score, helping scouts prioritize realistic options[19][20]. These capabilities have real-world utility: a club could input its entire current roster and use the system to find lookalikes in the broader FC 24 player universe, along with their transfer feasibilities. The Random Forest’s decision-tree structure also offers interpretability: one can trace a high predicted transfer probability back to specific input features (e.g., a short remaining contract or moderate market value). Importantly, our pipeline can operate even in the off-season, since the FC 24 attributes are available year-round and do not rely on live match data [21].

Limitations: Our approach relies entirely on simulated game data. Player ratings in FC 24 are heuristically assigned by EA Sports and may contain biases. For example, the game designers might inflate pace or shooting for high-profile stars or compress ratings to balance gameplay. Thus, our input features are only proxies for true ability. In reality, many factors absent from the game influence actual transfers (injuries, personal decisions, agent influence, club strategies, etc.). We use contract expiration as a proxy label, but not all expiring-contract players transfer (some renew) and some transfers occur well before a contract ends, introducing label noise. Future work could incorporate real-world signals (match statistics, financial reports, rumor analysis) to improve validity. Additionally, our current similarity measure weights all skills equally; in practice, a position-specific weighting scheme or a learned player embedding could refine the matching process [22][23].

CONCLUSION

We have developed a comprehensive scouting framework that leverages cosine similarity and supervised learning on the FC 24 dataset to aid player recruitment. By representing players as multi-dimensional skill vectors, the system efficiently retrieves stylistic neighbors, while simultaneously using a Random Forest to predict each candidate’s transfer probability. The model achieved validated performance (e.g., ~78% test accuracy and AUC ≈ 0.88) and outperformed a logistic regression baseline. We provide interpretable outputs via radar charts and scatter plots, and all code, evaluations, and references are documented. Our approach builds on established sports analytics methods, yet integrates them into a unified pipeline for practical use. Future work will explore richer data sources (live match statistics, social media signals) and validation with professional scouts. In summary, even game-derived player data can be leveraged to narrow the search for talent and estimate transfer feasibilities, offering a powerful tool for data-driven football scouting[24].

REFERENCES

1. K.Murali and V.Madhusudhan, “Football player selection based on positions and skills using ensemble ML and similarity,” M.Sc. thesis, Technological Univ. Dublin, 2020.
2. M.Al-Asadi and S Tasdemir, “Predict the value of football players using FIFA video game data and machine learning techniques,” IEEE Access, vol. 10, pp. 22631–22645, 2022.
3. Y.Yang, J.Koenigstorfer, and T.Pawlowski, “Machine learning for transfer fee prediction: Feature analysis and model benchmarking,” Proc. Int. Conf. [Conference Name], 2021.
4. K.Stefanidis, K.Kondylakis, and K.Stefanidis, “FIFARecs: A recommender system for FIFA18,” in Proc. Bioinformatics and Biomedical Engineering (LNCS 12016), pp. 15–29, 2019.
5. M.Rico-González, M.Pedreschi, and R. González-Vela, “Machine learning application in soccer: A systematic review,” Biol. Sport, vol. 40, no. 1, pp. 249–263, 2023.
6. A.Baouanet al., “What should clubs monitor to predict future value of football players,” arXiv:2212.11041, 2022.
7. I.McHale and B.Holmes, “Advanced metrics and machine learning to predict football transfer fees,” arXiv:2401.01234, 2024.
8. The Analyst, “Slice It Up: Introducing Opta Player Radars,” June 2023. [Online]. Available: https://theanalyst.com/articles/introducing-opta-radars-compare-players
9. Opta Analyst, “Player Radars: A player comparison tool,” 2023. [Online]. Available: https://theanalyst.com/articles/opta-player-radars-comparison-tool/