Towards Smarter Player Scouting: Learning Football Player Embeddings with Variational Autoencoders (VAEs)

Giulio Fantuzzi[1], Leonardo Egidi[1], Nicola Torelli[1]
[1] *University of Trieste*

Abstract

In this work, we propose the use of Variational Autoencoders (VAEs) to generate compact embeddings of football players. We investigate whether the learned latent space can capture key performance characteristics, enabling comparisons and clustering of players based on their playing style. These data-driven representations have the potential to enhance scouting processes by facilitating the identification of similar player profiles and uncovering patterns that may not be evident through traditional analysis. We evaluate our approach through preliminary experiments, showcasing VAEs's potential in extracting informative and robust player embeddings for advanced performance analysis.

Keywords: Variational Autoencoders (VAEs), Representation Learning, Football Player Embeddings, Data-Driven Scouting

1 Introduction

The growing availability of detailed statistical data from football matches has created new opportunities for advanced player analysis. Performance metrics describing passes, shots, and defensive actions provide valuable insights, yet their high dimensionality and sparsity pose challenges in extracting meaningful patterns. Representing players directly with raw statistics often leads to redundancy, noise, and difficulty in interpretation. To address this, we propose leveraging Variational Autoencoders [1] (VAEs) to learn a low-dimensional latent space that captures key performance characteristics. Unlike traditional dimensionality reduction methods (e.g., *PCA*, *MDS*, etc.), VAEs offer a probabilistic framework that structures the learned embeddings, enabling player clustering and similarity analysis based on underlying playing styles. Unlike standard Autoencoders (AEs), VAEs impose a probabilistic structure on the latent space, ensuring smoother and more meaningful representations that generalize better across different player profiles. The structure of this paper is as follows. Section 2 outlines the data collection and preprocessing steps, detailing the data sources and the feature engineering process. Section 3 introduces the framework of Variational Autoencoders (VAEs), discussing both its theoretical foundations and the specific configuration used in our study. In Section 4, we present and analyze the results of our preliminary experiments.

2 Data Collection and Preprocessing Steps

Due to the lack of a readily available dataset for our analysis, we built our own by integrating data from three different sources: `FBref` [2], `Transfermarkt`[3] and `clubelo`[4]. Our analysis focuses on the 2023/2024 season, with each source providing complementary information to enable a comprehensive evaluation.

FBref We extracted a wide range of player statistics covering various aspects of performance. These included shooting, passing, possession, defensive actions, goal-creating actions, and other miscellaneous statistics. Data was gathered for players across the following leagues: *Premier League, Championship, Serie A, Serie B, La Liga, La Liga 2, Bundesliga, Bundesliga 2, Ligue 1, Ligue 2, Eredivisie, Belgian Pro League, Primeira Liga, Major League Soccer, Liga MX*, and *Campeonato Brasileiro*, resulting in an initial dataset of 9,715 players.

Transfermarkt We collected biographical and contractual information, such as player position, age, height, preferred foot, and wage details.

clubelo We integrated club and league Elo ratings, which provide a measure of team ranking and competition strength. Since `clubelo` only offers data for European leagues, players from *MLS, Liga MX*, and *Campeonato Brasileiro* were excluded.

The data preprocessing involved several important steps. First, goalkeepers were excluded due to significant differences in their statistical profiles compared to outfield players. Next, players who had played fewer than 5 matches (intended as 450 minutes) were excluded to ensure meaningful statistical representation, leaving 6,152 players in the dataset. Feature engineering was performed by removing redundant statistics (e.g., *Goals-per-Shot*, *Tackles win percentage*, etc.) and converting all per-game metrics into per-90-minute statistics (P90). This standardization enabled more accurate comparisons between players. Data merging involved integrating `FBref` data with `Transfermarkt` data, which required resolving inconsistencies in player and club names across the two sources. This was addressed by implementing an automated matching algorithm based on the *Jaro-Winkler* [5] text similarity metric, with a confidence threshold of 0.9 set to prioritize matching accuracy over quantity. This resulted in a dataset of 5,639 players with complete information. After filtering for players with available `clubelo` rankings, the final dataset contains 4,541 players with 106 features, 97 of which were utilized to train the VAE. A more detailed description of the data and the scrapers used to collect it can be found in our GitHub repository.

3 Proposed Methodology

3.1 Latent Space Modeling: from Standard to Variational Autoencoders

Autoencoders (AEs) are models designed to learn a compressed representation of input data through an encoder-decoder structure. The encoder maps the input $\mathbf{x} \in \mathbb{R}^D$ to a lower-

dimensional representation $\mathbf{z} \in \mathbb{R}^d$, while the decoder provides a reconstruction $\tilde{\mathbf{x}} \in \mathbb{R}^D$ such that a reconstruction loss $\mathscr{L}(\mathbf{x}, \tilde{\mathbf{x}}) = ||\mathbf{x} - \tilde{\mathbf{x}}||^2$ is minimized. However, AEs do not impose any structure on the latent space, potentially leading to irregular and discontinuous embeddings, which may limit their usefulness for downstream tasks such as clustering and similarity analysis. VAEs address this limitation by introducing a *probabilistic framework*, modeling the latent variables as distributions rather than fixed points. Instead of learning a deterministic mapping $\mathbf{x} \mapsto \mathbf{z}$, the encoder approximates a posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ over the latent space, while the decoder generates a reconstruction by sampling $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ and mapping it back to the data space via $p_\theta(\mathbf{x}|\mathbf{z})$. The learning objective is to minimize the discrepancy between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the true (unknown) posterior $p(\mathbf{z}|\mathbf{x})$, which can be done by minimizing the following *Kullback-Leibler* (KL) divergence:

$$\begin{aligned} KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log q_\phi(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}) - \log p_\theta(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{x})] \\ &= \log p(\mathbf{x}) - \underbrace{\Big(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}))\Big)}_{\text{ELBO } \mathscr{L}(\phi,\theta,\mathbf{x})}. \end{aligned} \tag{1}$$

Minimizing the KL divergence with respect to parameters ϕ and θ is equivalent to maximizing the Evidence Lower Bound (ELBO) $\mathscr{L}(\phi, \theta, \mathbf{x})$. Its first term is a reconstruction loss, which ensures that the decoder learns to generate realistic data from the latent space, while the KL term regularizes the latent space by encouraging it to follow a prior distribution $p(\mathbf{z})$.

Gaussian Assumption In our implementation, we assume a Gaussian form for both the approximate posterior and the prior. This choice is standard in VAEs because (1) assuming a Gaussian likelihood, the reconstruction term corresponds to a mean squared error loss, and (2) the KL divergence has a closed form expression, simplifying optimization.

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathscr{N}(\mathbf{z}; \mu(\mathbf{x}; \phi), \sigma^2(\mathbf{x}; \phi)\mathbf{I}), \tag{2}$$

$$p(\mathbf{z}) = \mathscr{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

$$KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})) = \frac{1}{2} \sum_{i=1}^{d} \left[\sigma_i^2(\mathbf{x}; \phi) + \mu_i^2(\mathbf{x}; \phi) - 1 - \log \sigma_i^2(\mathbf{x}; \phi)\right]. \tag{4}$$

Reparameterization Trick Sampling the latent variables $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x})$ after the encoding is a non-differentiable operation, which is an issue when we want to use backpropagation and make gradient descent possible on this architecture. We hence adopted the so called *reparametrization trick*, which solves this issue expressing $\mathbf{z}$ as follows:

$$\mathbf{z} = \mu(\mathbf{x}; \phi) + \sigma(\mathbf{x}; \phi) \odot \epsilon, \quad \epsilon \sim \mathscr{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

allowing gradients to flow through μ and σ and enabling efficient end-to-end training.

3.2 VAE Architecture and Training Configuration

The VAE used in this work consists of an encoder with two fully connected layers that progressively reduce the 97-dimensional input representation to 64 and then 32, applying `ReLU` activations at each step. The latent space has a dimension of 16, and the decoder mirrors the encoder structure, reconstructing the original 97-dimensional feature space. The model was implemented in `PyTorch` and trained using the `AdamW` [6] optimizer.

The training configuration is summarized in Table 1.

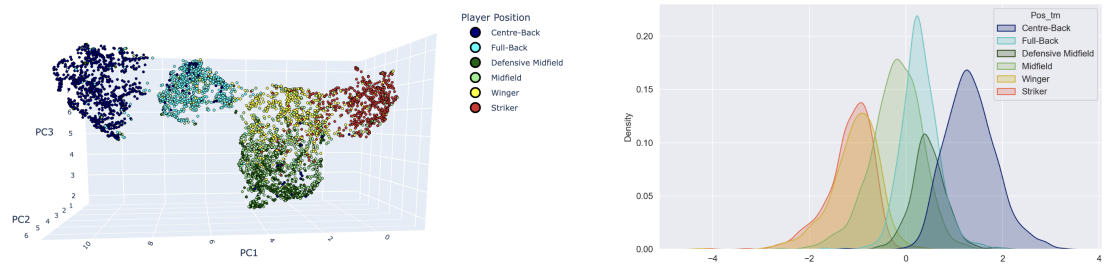
Table 1: Training configuration

Hyperparameter	**Value**
Learning Rate	$1e-2$
Betas	(0.9, 0.99)
Weight Decay	$1e-3$
Batch Size	512
Training Epochs	1200
Weight Initialization	Kaiming
Learning Rate Scheduler	`ReduceLROnPlateau`

4 Experimental Results and Analysis

4.1 Latent Space Visualization

To gain insights into the structure of the latent space, we applied our VAE's encoder to the data, using the means of the latent distribution as player embeddings. To visualize them, we employed UMAP [7] to reduce the dimensionality and plot the first three principal components. This approach reveals a well-structured latent space, where meaningful patterns emerge in the distribution of data points, as shown in Figure 1a. Centre-backs and full-backs are grouped in distinct, standalone clusters, indicating that the model has successfully captured and separated these positions based on their unique playing styles. Full-backs are located near wingers, which is interpretable since both typically operate on the sides of the pitch and share similar tactical roles. Midfielders exhibit some clustering, though less distinct. Among them, defensive midfielders are positioned farther from wingers, highlighting a key difference in their playing characteristics. Strikers form a less compact group compared to defenders, but are still clearly distinguishable. Interestingly, wingers act as a bridge between midfielders and strikers, reflecting their dual role in both attack and midfield support. The latent space organization is particularly compelling, as the model has learned a structured representation of player roles and playing styles despite the absence of such information during training. This suggests that our VAE captures relationships purely through performance characteristics, uncovering latent patterns tied to on-field behavior rather than predefined positional labels. Supporting this observation, Figure 1b displays the distribution of a selected latent component, where distinct density curves for different player roles further confirm that the learned space inherently encodes positional differences.



(a) UMAP visualization of latent space.

(b) Latent component analysis.

Figure 1: Latent space visualization.

4.2 Player Similarity Analysis

To further assess the quality and utility of the learned embeddings, we conducted a player similarity analysis. Players with closer embeddings, measured by Euclidean distance, are considered more similar. Table 2 presents selected examples of well-known footballers alongside their most similar counterparts based on the learned representations. The first example focuses on *Rafael Leao*, a dynamic winger renowned for his speed and dribbling ability. Notably, despite the model being position-agnostic, all the identified similar players are also wingers. Moreover, football enthusiasts will recognize that the listed players - such as *Nico Williams*, *Marcus Edwards* and *Mason Greenwood*- are indeed fast, technical attackers who excel in one-on-one situations, reinforcing the validity of the similarity analysis. Next, *Olivier Giroud*'s case introduces a more nuanced example. Most of the retrieved players share his profile as a physically dominant striker, proficient in hold-up play and finishing. The inclusion of *Haris Tabakovic* and *Che Adams*, who were playing in lower-tier leagues (*Bundesliga 2* and *Championship*, respectively), highlights an important feature of the model: while league and team rankings were included in the model, they do not overly bias the embeddings, allowing meaningful comparisons across different competition levels. For *Francesco Acerbi*, the model identifies centre-backs with comparable profiles. While the insights in this case may be less striking, the results demonstrate the model's effectiveness across defensive roles. Lastly, *Jamal Musiala*'s similarity to emerging talents like *Bradley Barcola* and *Lamine Yamal* suggests that the model is capable of identifying rising stars, underscoring its potential for scouting and talent identification.

5 Conclusions

In this work, we explored the application of Variational Autoencoders (VAEs) to generate compact and informative embeddings of football players. By leveraging a probabilistic framework, we were able to learn a structured latent space that captures essential performance characteristics, forming a solid foundation for advanced player analysis. The embeddings generated by our model facilitate player comparison, clustering, and scouting, opening new avenues for identifying similar player profiles and uncovering latent patterns that may not

Table 2: Scouting application: comparison of player profiles based on embedding similarity

Rank	Player Name	Position	Age	Height (cm)	Preferred Foot	Squad	League
	Rafael Leao	**Winger**	**24**	**188**	**Right**	**Milan**	**Serie A**
1	Nico Williams	Winger	21	181	Right	Athletic Club	La Liga
2	Marcus Edwards	Winger	24	168	Left	Sporting CP	Primeira Liga
3	David Neres	Winger	26	176	Left	Benfica	Primeira Liga
4	Edon Zhegrova	Winger	24	180	Left	Lille	Ligue 1
5	Mason Greenwood	Winger	21	181	Left	Getafe	La Liga
	Olivier Giroud	**Striker**	**36**	**192**	**Left**	**Milan**	**Serie A**
1	Niclas Fullkrug	Striker	30	189	Right	Dortmund	Bundesliga
2	Haris Tabakovic	Striker	29	194	Right	Hertha BSC	Bundesliga 2
3	Jonas Wind	Striker	24	190	Right	Wolfsburg	Bundesliga
4	Che Adams	Striker	27	179	Right	Southampton	Championship
5	Marko Arnautovic	Striker	34	190	Left	Inter	Serie A
	Francesco Acerbi	**Centre-Back**	**35**	**192**	**Left**	**Inter**	**Serie A**
1	Nicolo Casale	Centre-Back	25	191	Right	Lazio	Serie A
2	Antonio Rudiger	Centre-Back	30	190	Right	Real Madrid	La Liga
3	Lukas Klostermann	Centre-Back	27	189	Right	RB Leipzig	Bundesliga
4	Hauke Wahl	Centre-Back	29	189	Right	St. Pauli	Bundesliga 2
5	Gianluca Mancini	Centre-Back	27	190	Right	Roma	Serie A
	Jamal Musiala	**Midfield**	**20**	**184**	**Right**	**Bayern Munich**	**Bundesliga**
1	Bradley Barcola	Winger	20	182	Right	Paris S-G	Ligue 1
2	Brahim Diaz	Winger	23	170	Left	Real Madrid	La Liga
3	Amine Adli	Winger	23	174	Left	Leverkusen	Bundesliga
4	Lamine Yamal	Winger	16	180	Left	Barcelona	La Liga
5	Ibrahim Salah	Winger	21	186	Right	Rennes	Ligue 1

be immediately visible through traditional statistical metrics. Future research will combine player statistics with tracking and event data to capture action-based information, player interactions, and tactical movements. It will also investigate feature importance and methods to condition the similarity measure, optimizing player replacements for specific team needs.

References

[1] Diederik Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 12 2013.

[2] FBref. `https://fbref.com/en/`. Accessed: 2025-01-15.

[3] Transfermarkt. `https://www.transfermarkt.com`. Accessed: 2025-01-15.

[4] ClubElo. `http://clubelo.com`. Accessed: 2025-01-15.

[5] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida, 1989.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[7] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.