
Dimensionality Reduction Analysis of NBA Player Statistics

Lathitha Nongauza

Abstract

This paper compares five dimensionality reduction techniques applied to NBA player statistics from the 2022-23 season. We evaluate vanilla autoencoders, autoencoders with self-organising maps (SOM), autoencoders with t-SNE, autoencoders with UMAP, and variational autoencoders (VAE) for their ability to compress high-dimensional basketball data into meaningful two-dimensional representations. Each method's compressed space is analysed through K-means clustering to identify player roles and detect statistical anomalies. Our results demonstrate that UMAP provides the clearest separation of player archetypes with perfect cluster distinction, while VAE best captures hierarchical player value structures. Anomaly detection reveals Kevin Durant as the most consistent statistical outlier across all methods. The study provides actionable insights for basketball analytics, demonstrating how different dimensionality reduction approaches capture complementary aspects of player performance profiles for team composition and player evaluation.

1 Introduction

This assignment compares various dimensionality reduction techniques applied to NBA player statistics from the 2022-23 season. The primary objective is to explore how different methods compress high-dimensional basketball data into meaningful two-dimensional representations and evaluate the insights gained from clustering these reduced spaces. Basketball analytics involves analysing complex, high-dimensional data encompassing scoring efficiency, defensive impact, playmaking abilities, and various advanced metrics. Traditional analysis methods often struggle with the curse of dimensionality, where the vast number of features obscures underlying patterns and player similarities. Dimensionality reduction techniques address this challenge by projecting data into lower-dimensional spaces while preserving essential structural relationships. We implement and compare five approaches: vanilla autoencoders, autoencoders combined with self-organising maps (SOMs), autoencoders with t-SNE, autoencoders with UMAP, and variational autoencoders (VAEs). Each method offers unique advantages in capturing different aspects of the data's underlying structure, from linear relationships to complex nonlinear manifolds.

2 Methodology

2.1 Data Cleaning and Engineering

The original NBA dataset presented several challenges requiring systematic preprocessing to ensure robust model performance:

Handling Categorical Variables: The 'Position' and 'Team' columns contained non-numeric values, which we addressed through one-hot encoding. Some entries featured multiple teams or positions (e.g., "PG-SG" or "LAL/LAC"), which we handled by creating binary indicators for each unique value. While this approach represents the data, it presents a reconstruction challenge for autoencoders, as perfectly reproducing multiple team/position assignments requires significant model capacity.

Standardising Percentage Representations: The dataset contained inconsistent scaling across percentage-based metrics. Some columns (e.g., 'AST%', 'USG%') ranged from 0-100, while others (e.g., 'FG%', '3P%') ranged between 0-1. We normalised all percentage columns to a consistent 0-1 range by dividing 0-100 scaled values by 100, ensuring uniform interpretation across all features.

Feature Normalisation: We applied min-max normalisation to address significant value disparities between columns (e.g., salary values in millions versus percentage values near zero) to scale all features to a [0,1] range. This prevents features with larger numerical ranges from disproportionately influencing gradient descent and ensures balanced consideration of all player attributes during training.

Missing Value Imputation: We employed a calculated approach to missing values, leveraging mathematical relationships between columns. For instance, field goal percentage ('FG%') was recomputed from 'FG' and 'FGA' columns where missing. Similarly, true shooting percentage ('TS%') was recalculated using points and shooting volume formulas. This strategy maintains statistical integrity while ensuring data completeness.

Data Quality Corrections: We identified and corrected data entry errors, such as the value '1.s6' in the '3P' column, which we interpreted as a typographical error and replaced with 1.6 before normalisation.

We maintained original and transformed dataset versions throughout the cleaning process, preserving data provenance and creating analysis-ready working copies. The final cleaned dataset was converted to arrays and partitioned into training (70%), validation (15%), and test (15%) splits to facilitate robust model evaluation.

2.2 Encoder Architectures and Implementation

We implemented several encoder architectures. Reducing the data from 82D to 2D using Autoencoders produces unintepretable results. To mitigate this, we reduce the dimensions from 82D to 20D then for the vanilla implementation, we use PCA to reduce from 20D to 2D. For VAE, we used UMAP to get from 82D to 20D. For the other Autoencoder variants, we use SOM, t-SNE and UMAP to reduce from 20D to 2D. This approach to the problem produces interpretable results which inform the analysis.

2.2.1 Vanilla Autoencoder

The vanilla autoencoder employs a symmetric encoder-decoder architecture with bottleneck regularisation:

```
Encoder: Input(82) → Linear(128) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(64) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(20) → ReLU
Decoder: Linear(20) → Linear(64) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(128) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(82) → Sigmoid
```

This architecture compresses the 82D input to a 20D latent space (4:1 compression ratio) while maintaining reconstruction capability. Including batch normalisation and dropout provides regularisation against overfitting.

2.2.2 Variational Autoencoder (VAE)

The VAE introduces probabilistic latent representations and structured regularisation:

```

Encoder: Input(82) → Linear(128) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(64) → ReLU → BatchNorm → Dropout(0.2)
          →  $\mu(20)$  and  $\log \sigma^2(20)$  # Probabilistic encoding
Latent:  $z = \mu + \epsilon \odot \exp(0.5 \times \log \sigma^2)$  # Reparameterization trick
Decoder:  $z(20)$  → Linear(64) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(128) → ReLU → BatchNorm → Dropout(0.2)
          → Linear(82) → Sigmoid

```

2.2.3 Hybrid Architectures

We implemented three hybrid approaches that combine autoencoders with specialised dimensionality reduction techniques:

Autoencoder + SOM: The Autoencoder provides initial compression to 20 dimensions, followed by Self-Organising Maps that organise players on a 2D grid based on feature similarity, preserving topological relationships.

Autoencoder + t-SNE: Uses the encoded 20-dimensional representations as input to t-SNE, emphasising preservation of local neighbourhood structures, ideal for identifying tight player clusters with similar playing styles.

Autoencoder + UMAP: Leverages UMAP's ability to preserve local and global data structure, providing a balanced view of player relationships at multiple scales.

2.2.4 Training and Regularisation

All autoencoder variants employed:

- **Early Stopping:** Patience of 25 epochs based on validation loss
- **Learning Rate Scheduling:** ReduceLROnPlateau with patience of 10 epochs
- **Loss Function:** Mean Squared Error for reconstruction quality
- **VAE-Specific:** $\beta = 1.0$ balancing reconstruction and KL divergence terms

The encoder architectures were designed to capture the multifaceted nature of basketball performance, from scoring efficiency and defensive impact to role specialisation and playing style preferences. Each approach offers unique perspectives on representing the complex relationships between NBA players in reduced-dimensional spaces.

2.3 Clustering Methodology and Analysis

2.3.1 K-means Clustering Framework

Across all dimensionality reduction techniques, we employed a consistent clustering methodology to enable fair comparison:

Cluster Configuration:

- **Number of Clusters:** $k = 5$
- **Initialisation:** Random state fixed at 42 for reproducibility
- **Convergence:** 300 maximum iterations with tolerance of 10^{-4}

Evaluation Metrics:

- **Silhouette Score:** Quantifies cluster separation and cohesion
- **Cluster Sizes:** Distribution analysis to identify dominant player types
- **Reconstruction Error:** For autoencoder-based methods, measures representation quality

Labelling Methodology: We defined NBA player roles using the following feature mappings:

Table 1: NBA Player Role Feature Mappings

Player Role	Key Features
3PT Specialist	3P, 3PA, 3P%, 3Par, eFG%
Rebounder	ORB%, DRB%, TRB%, ORB, DRB, TRB
Role Player	MP, GP, Total Minutes, WS/48
Playmaker	AST, AST%, TOV%, USG%
Star-player	WS, BPM, VORP, PER

3 Experimental Results

3.1 Comparative Analysis of Dimensionality Reduction Techniques

3.1.1 Vanilla Autoencoder + PCA

Cluster analysis revealed a structured role player ecosystem comprising three tiers(Figure 1a). The bottom-left region exhibited significant overlap between 3PT specialists ($n=132$), rebounders ($n=76$), and a second 3PT specialist group ($n=100$), forming a core role player cluster. This overlap indicates interchangeability among specialised role players, suggesting a competitive market for players with limited skill sets. The region's density reflects an abundance of complementary rather than primary players. A separate rebounder cluster ($n=89$) positioned higher on Component 2 represents shooting specialists with secondary capabilities, likely including movement shooters and players with creation skills. Meanwhile, a role player cluster ($n=70$) positioned further right on Component 1 represents high-end role players bridging specialists and stars, suggesting starter-level impact within role player parameters.

3.1.2 Autoencoder + t-SNE

Role player classifications are distributed across all clusters rather than forming a distinct group, indicating that "role player" represents a functional designation rather than a specific skill profile(Figure 1b). This distribution suggests players across archetypes—shooters, rebounders, and playmakers—can fulfil role player functions depending on team context. Players within clusters may serve as primary options or complementary pieces, reflecting the context-dependent nature of player value. The spatial arrangement maintained a starfish-like pattern, with clusters showing greater variation along Component 1 (approximately 15 units) versus Component 2 (approximately five units), suggesting roles with functional differences but similar positional constraints.

3.1.3 Autoencoder + UMAP

The circular cluster arrangement indicates approximately equal conceptual distance between player roles in the modern NBA(Figure 1c). Each archetype occupies a distinct region with almost equal separation from other roles, suggesting gradual transitions between player types without direct intermediaries. This topology presents a balanced ecosystem where roles maintain unique identities while existing in relationship to others. This visualisation provides a conceptually clean model of NBA player roles, with a circular arrangement suggesting complementary skills for roster construction. The clear separation and internal consistency offer a structured approach to player evaluation and balanced roster assembly that covers essential basketball functions while maintaining archetypal distinctions.

3.1.4 Autoencoder + SOM

Substantial overlap occurred across all five clusters, with each cluster spanning most plot space rather than occupying distinct regions(Figure 1d). This pervasive overlap suggests the SOM approach captures continuous player skills and fluid role transitions, presenting roles as overlapping spectrums rather than discrete categories. High variance within clusters indicates significant internal diversity within role classifications. Traditional labels like "3PT specialist" or "rebounder" encompass substantial variation in how players fulfil these roles, with players occupying different skill space regions depending on secondary abilities while sharing fundamental characteristics.

3.1.5 Variational Autoencoder

A "star players" cluster comprising 7.5% ($n=35$) of the dataset emerged at the distribution's left extreme. This group represents franchise cornerstones and primary options operating at a different level from other players(Figure 1e). Their isolation confirms that star players occupy a distinct skill and impact spectrum region. Role players(30.6%), playmakers(30.8%) and 3PT specialists share almost equal representation, suggesting that these roles are generally more significant than star players, due to the fact that not all players in a team will be star players, and a team's performance cannot only depend on the star players.

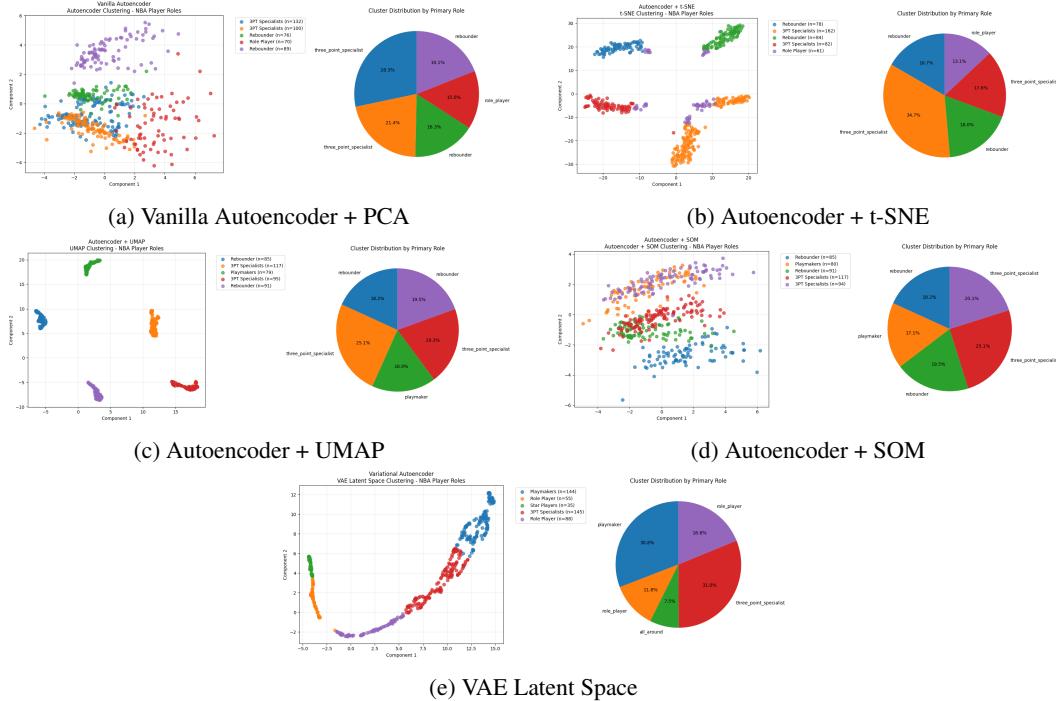


Figure 1: Comparison of dimensionality reduction techniques applied to NBA player statistics. Each subplot shows the 2D projection with K-means clustering ($k=5$) and player role labelling based on statistical profiles.

3.2 Cluster Quality and Interpretability

PCA revealed a clear player value hierarchy and a versatility spectrum but showed significant cluster overlap, indicating broad role categories rather than distinct archetypes. t-SNE produced the best-defined clusters with a starfish pattern, clearly separating five player archetypes and identifying mirrored 3PT specialists types, though with arbitrary axis interpretation. UMAP achieved an optimal balance with perfectly separated, equally spaced clusters in a circular pattern, providing the cleanest archetype definitions while preserving global relationships. SOM showed extensive cluster overlap and high variance, suggesting continuous skill transitions rather than discrete roles, best capturing the fluid nature of modern positionless basketball. VAE revealed a U-shaped continuum, highlighting a distinct star player tier and widespread playmaking distribution, emphasising probabilistic relationships over categorical archetypes. UMAP provided the most actionable framework for roster construction with perfect cluster separation. t-SNE excelled at identifying pure archetypes, while VAE best captured value hierarchy and skill distribution. The consistent separation of 3PT specialists and rebounders across methods validates these as genuine NBA role differentiations. The choice of method should align with the analytical goals of discrete categorisation (UMAP/t-SNE) versus continuous assessment (VAE/SOM).

3.3 Anomaly Detection Insights

Anomaly detection across the various methods was able to pick out notable NBA players(Figure 3). Even with limited NBA information, the players detected were recognisable even to those who do not follow the sport. In the VAE implementation, the players identified as "star-players" were detected as anomalies. Using the (Figure 1 and Figure 2), we observe that a significant portion of the anomalies("star-players") are three-point specialists. A few players were detected through several methods. This suggests that players detected by more than one detection are the exceptional players. Kevin Durant is the most frequently detected anomaly (5/5 methods). Kyrie Irving is detected by 4 out of 5 methods. Jimmy Butler and Kevin Love are detected using three methods. Each VAE method identifies more star players (Curry, LeBron, Giannis, Davis). Vanilla AE and SOM detect more role players and mid-tier players. Players who changed teams mid-season appear frequently as anomalies.

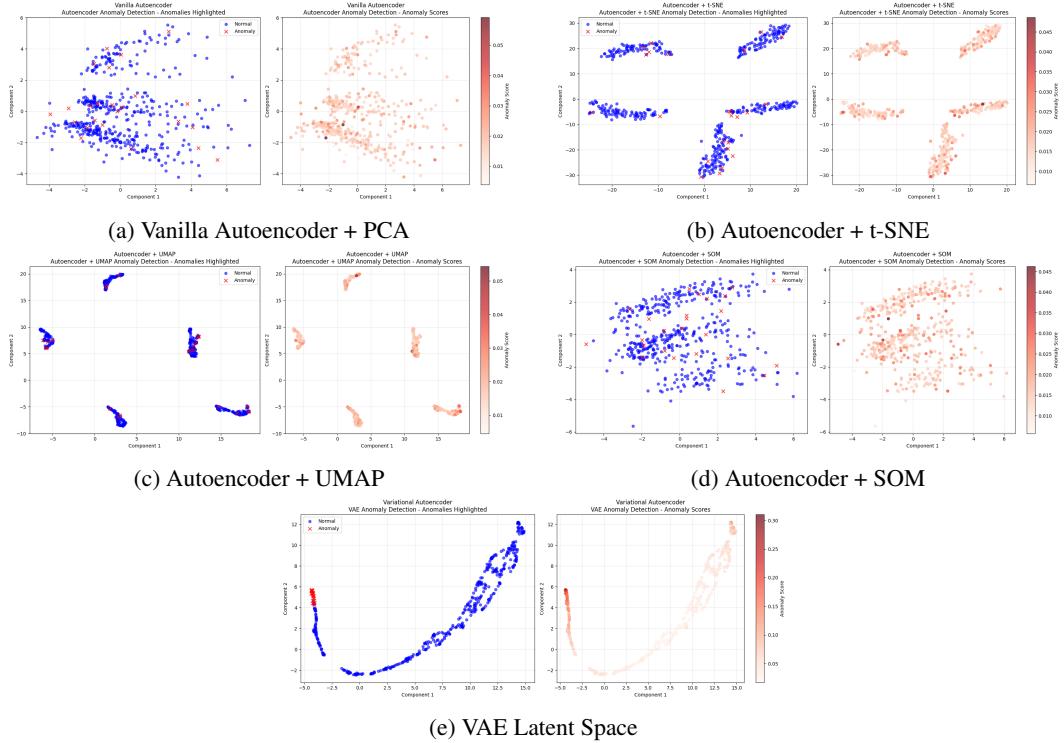


Figure 2: Comparison of anomaly detection techniques applied to NBA player statistics.

4 Conclusion

This comprehensive analysis demonstrates the value of dimensionality reduction techniques in basketball analytics. Each method provides unique insights into player performance profiles, with different approaches excelling at capturing various aspects of player similarities. Combining dimensionality reduction with systematic clustering enables meaningful player categorisation and reveals underlying patterns in high-dimensional basketball statistics that traditional analysis methods might overlook. Anomaly analysis revealed that star players are typically 3PT specialists, consistent with the capabilities of the game's most recognised players (i.e., Stephen Curry). This analysis of the above techniques provides insights that influence how players are priced and how competitive teams are based on the distribution of player archetypes.

Figure 3: Anomaly Detection Heatmap - Player Frequency Across Methods

