Generating Synthetic FIFA Player Attributes Using an IT-GAN-Inspired Framework

Manh Cuong La & Binh Minh Nguyen

UMassAmherst
College of Engineering
Electrical and Computer Engineering

Motivation

- Soccer teams rely heavily on player statistics for scouting and tactical planning
- Real player data is often private, limited, or not shareable due to licensing
- Researchers cannot easily test “what-if” scenarios (e.g., new types of players)
- Synthetic, realistic data would let teams and analysts explore player variations safely

Problem Statement

- Real soccer player statistics cannot be freely shared
- Generating realistic but privacy-safe synthetic players is needed
- GOAL: learn relationships between attributes (age, speed, passing, etc.) and sample new players that keep those patterns without copying any real player
- Create new player profiles that preserve statistical patterns but protect privacy by not reproducing any real player

Selected Paper Overview

Paper: *Invertible Tabular GANs: Killing Two Birds with One Stone for Tabular Data Synthesis (NeurIPS 2021)* [1]

- IT-GAN introduces a principled framework for tabular data synthesis
- Combines three components:
  - Autoencoder → learns compact latent representation
  - Invertible flow (NODE-based) → exact, bijective density modeling
  - WGAN-GP → adversarial training for realistic samples
- Provides a one-to-one mapping between latent and data space, enabling likelihood-based training and privacy–quality trade-offs

Dataset Overview

- Sourced from the *FIFA 24 Player Stats Dataset* (Kaggle) [2]
- Contains 5682 players with 23 attributes
- Includes technical, physical, and mental stats (e.g., passing, dribbling, strength)
- Data cleaned and preprocessed before training
- Selected 23 numerical attributes for modeling

Preprocessing Data

- Removed non-numeric fields (name, club, nationality)
- Kept only continuous gameplay attributes
- Handled missing values (mean imputation)
- Normalized features to [−1,1] for stable GAN and flow training
- Final dataset shape: 5682 players × 23 features

Model Architecture

- Simplified IT-GAN: Autoencoder + RealNVP flow + WGAN-GP critic
- Autoencoder: compresses player stats into a latent vector
- RealNVP Flow: enforces invertibility and structured latent distribution
- WGAN-GP Critic: trains generator to match real data distribution
- Allows smooth generation and stable density modeling

Training Setup

- Training performed in Google Colab (GPU accelerated)
- Epochs: 300 (losses stabilized after ~200 epochs)
- Batch size: 256
- Optimizers: Adam (lr = 2e-4 for AE & flow, lr = 1e-4 for critic)
- Losses:
  - Reconstruction loss (autoencoder)
  - Wasserstein loss + gradient penalty (critic)
  - Flow log-likelihood loss

Synthetic Player Generation

- Generate players by sampling random latent vectors
- Flow transforms latent noise into structured latent codes that decode into realistic player profiles
- Decoder converts latent representation into full attribute set
- Output: 2000 synthetic players
- Values rescaled back to real FIFA attribute ranges (0–100)

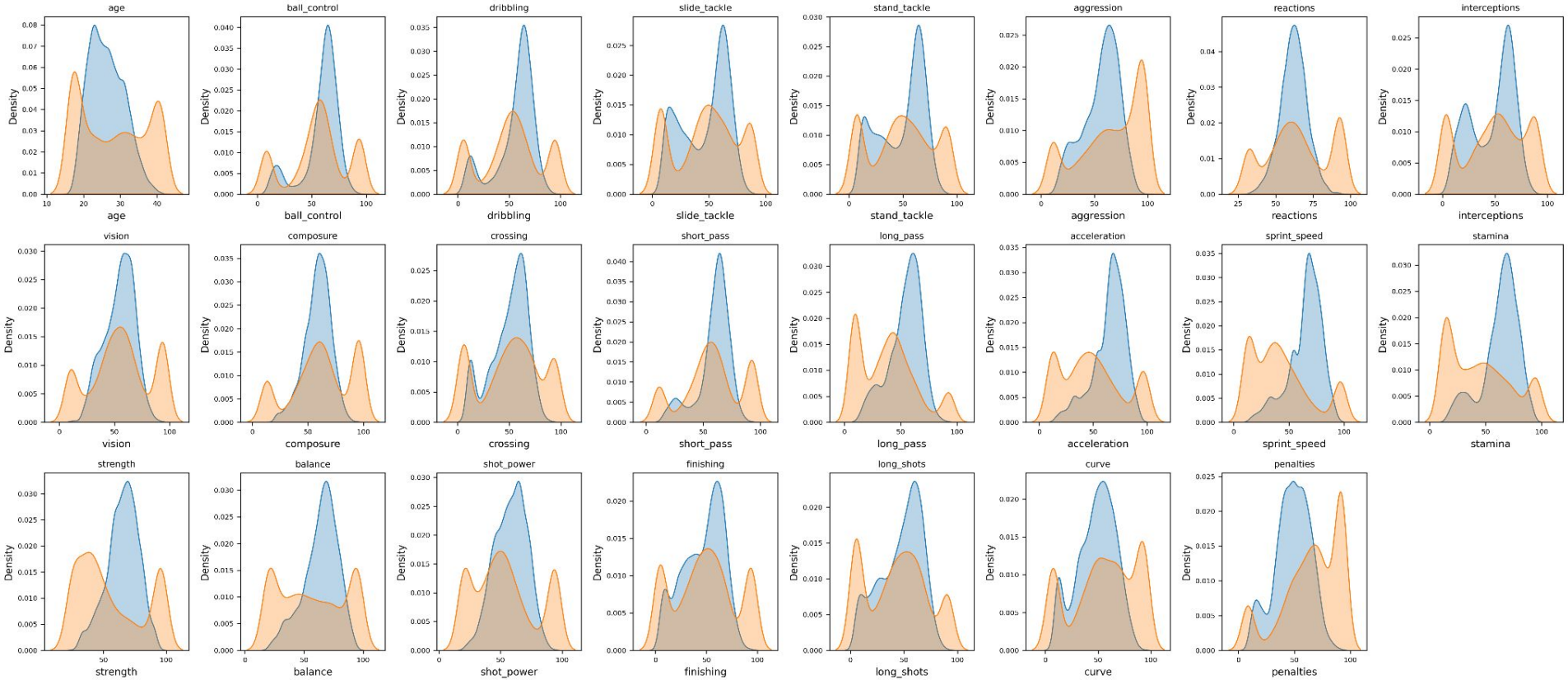
Evaluation Methods

- Distribution Matching: KDE curves (real vs synthetic)
- Summary Statistics: compare means and variances
- Wasserstein Distance per feature (lower = better similarity)
- Correlation Heatmaps: check if relationships between features are preserved
- PCA Visualization: shows real and fake players overlap in feature space

KDE Distributions

- Real (blue) vs. fake (orange) KDE curves for all 23 attributes
- Many technical skills (e.g., ball_control, finishing, vision) show very similar shapes, meaning the generator learned the overall trends
- Largest mismatches occur for physical stats (e.g., sprint_speed, stamina, acceleration) and some skills (e.g., penalties, long_pass)
- KDE plots let us visually check realism and quickly spot which attributes the model struggles to capture

KDE Distributions (cont.)



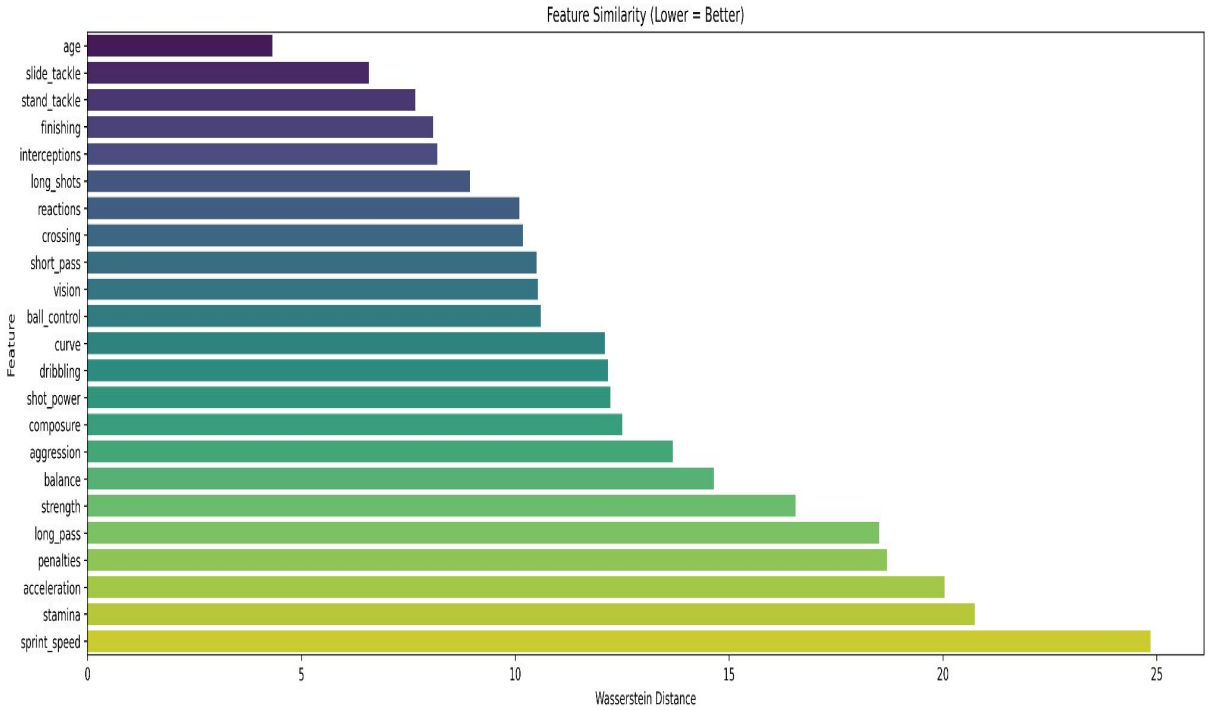
Summary Statistics (Mean & Std)

- Many features have very similar means between real and fake players *(e.g., ball_control, passing, finishing)*
- Largest differences appear in a few attributes:
  - age (fake players slightly older)
  - strength (fake slightly lower)
  - long_pass (fake significantly lower)
- Fake data has higher variability (std) across most attributes → The generator produces a wider range of players.
- Overall, synthetic players follow the real trends but are not exact copies - good for privacy and diversity

	Real Mean	Fake Mean	Real Std	Fake Std
age	26.320000	28.160000	4.730000	9.000000
ball_control	58.910000	55.930000	16.570000	26.500000
dribbling	56.130001	51.590000	18.770000	29.000000
slide_tackle	46.730000	47.830002	20.520000	27.209999
stand_tackle	48.820000	49.009998	20.980000	28.590000
aggression	56.320000	65.220001	16.850000	29.080000
reactions	61.959999	64.720001	8.890000	20.389999
interceptions	47.389999	48.169998	20.450001	29.820000
vision	54.470001	55.480000	13.710000	26.639999
composure	58.619999	61.880001	12.020000	26.700001
crossing	49.790001	51.290001	17.900000	29.480000
short_pass	59.330002	57.639999	14.330000	25.040001
long_pass	53.910000	38.349998	14.600000	24.790001
acceleration	64.750000	49.020000	15.300000	28.129999
sprint_speed	64.959999	43.119999	15.110000	26.570000
stamina	63.380001	45.419998	16.110001	27.540001
strength	65.379997	55.900002	12.620000	25.700001
balance	64.070000	56.790001	14.500000	27.540001
shot_power	58.180000	53.790001	12.970000	24.910000
finishing	46.349998	48.630001	19.820000	29.709999
long_shots	47.160000	43.470001	19.459999	27.980000
curve	48.099998	55.599998	18.090000	29.389999
penalties	48.169998	65.370003	15.780000	25.469999

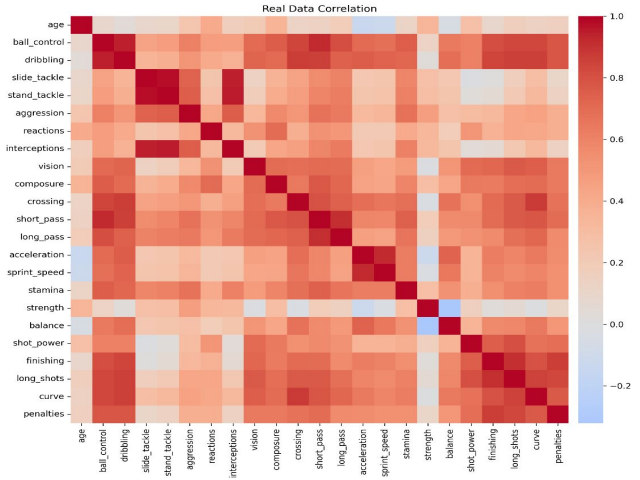
Wasserstein Distances

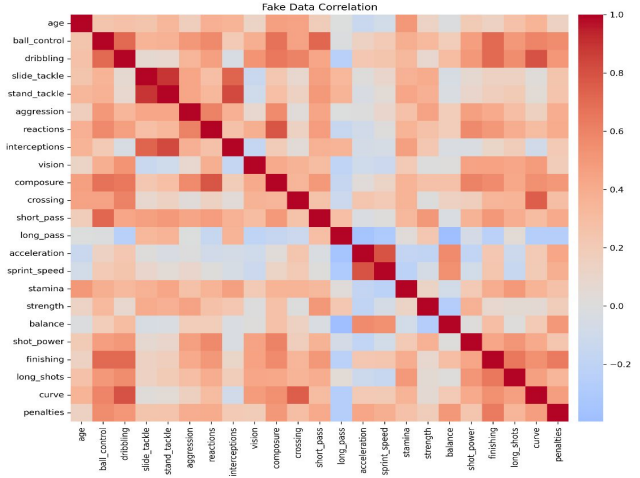
- Measures how different the real and synthetic feature distributions are
- Lower values indicate stronger similarity between real and generated players
- Most attributes show relatively small distances, which means the model captures general trends well
- A few attributes stand out with higher distances — especially sprint speed, stamina, acceleration, penalties, long_pass, and strength
- Results suggest overall good performance but room for improvement on certain attributes



Correlation Heatmaps

- Compare relationships between attributes in real vs. synthetic data: real data shows strong clusters (e.g., speed–acceleration–stamina, dribbling–ball control–vision, finishing–shot power)
- Fake data preserves many of these high-level correlation patterns
- Some correlations appear weaker, noisier, or missing in synthetic data. Especially interactions involving physical stats (speed, stamina) and some technical skills
- Indicates IT-GAN captures the overall structure but struggles with more complex multi-attribute dependencies
- Correlation heatmaps help verify whether the model learned interactions, not just individual distributions

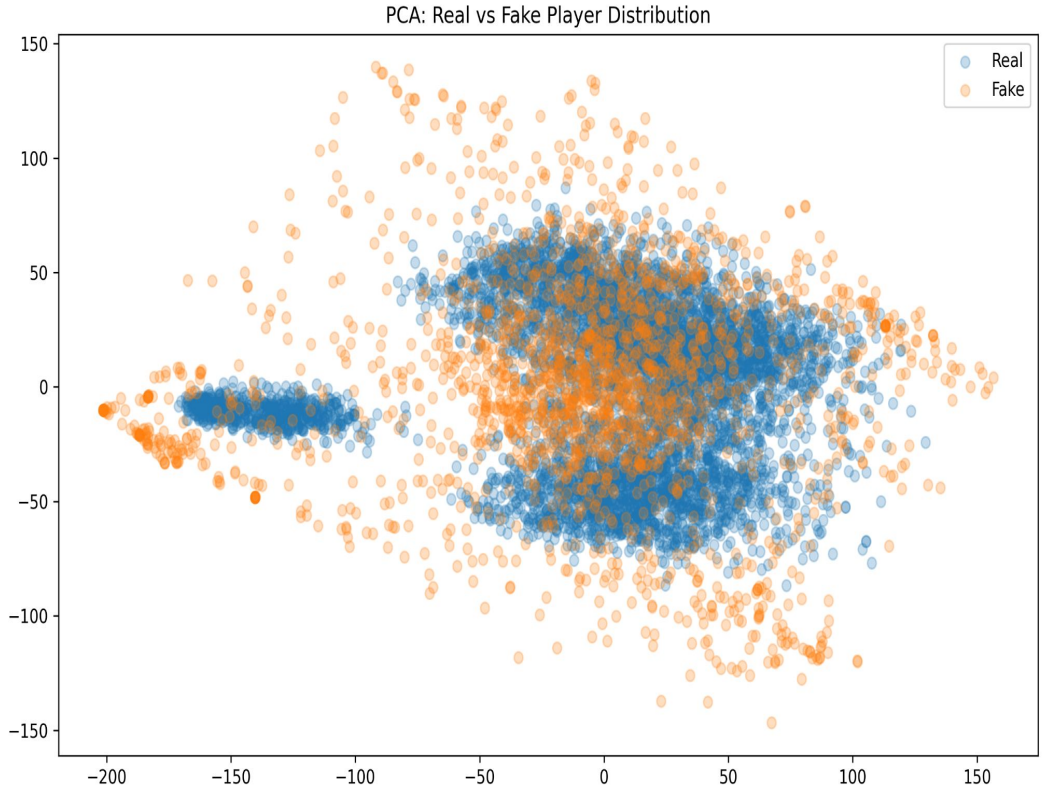




PCA Visualization

- PCA projects the 23-dimensional player attributes into 2D for easier comparison
- Real (blue) and fake (orange) players overlap heavily, showing the generator learned the broad player distribution
- Both real and synthetic data form two similar clusters (e.g., physical players vs. technical players)

  → Fake players correctly follow the real clustering structure
- Fake players show slightly larger spread and more outliers, indicating some variation is not fully captured
- Importantly, synthetic players do not form unnatural or separate clusters — a strong sign of realism



References

[1] J. Lee, J. Hyeong, J. Jeon, N. Park, and J. Cho,

"Invertible Tabular GANs: Killing Two Birds with One Stone for Tabular Data Synthesis," arXiv preprint arXiv:2202.03636, 2022.

https://arxiv.org/pdf/2202.03636

[2] Rehan D. Lodhi, "FIFA 24 Player Stats Dataset," Kaggle, 2023.

https://www.kaggle.com/datasets/rehandl23/fifa-24-player-stats-dataset

Questions?

University of Massachusetts Amherst