PROCEEDINGS OF SCIENCE

Transfer Learning under Domain Shift in Football Analytics

Luca Clissa[a,b,*] **and Antonio Macaluso**[c]

[a] *Department of Physics and Astronomy "Augusto Righi", University of Bologna*
[b] *Istituto Nazionale di Fisica Nucleare, Sezione di Bologna*
[c] *German Research Center for Artificial Intelligence (DFKI)*

E-mail: luca.clissa2@unibo.it, antonio.macaluso@dfki.de

Machine learning models trained on football data often degrade when transferred across leagues, seasons, gender categories, or tournament types. Rather than treating this as a nuisance, we argue that football is a particularly rich benchmark for transfer learning and domain adaptation, because it naturally exhibits prior shift, covariate shift, and concept shift at the same time. In this work we study action-based goal-scoring prediction under cross-competition shift using publicly available StatsBomb data and an open-source SPADL/VAEP processing pipeline. We represent each decision point with a three-action game state and train an XGBoost classifier to predict whether the team in possession will score within the next ten actions. The source model is pretrained on several seasons of La Liga men's competition. Generalization performance is then tested using as target domain other major European leagues, thus exposing season-, league- and combined league-season-shifts. Although calibration and classification remain challenging due to class imbalance, noisy labels, and competition-specific dynamics, preliminary evidence shows how fine-tuning provides small but consistent gains over zero-shot transfer, while target-only models catch up as more target labels become available.

International Symposium on Grids and Clouds (ISGC2026)
15-20, March, 2026
Academia Sinica, Taipei, Taiwan

*Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) All rights for text and data mining, AI training, and similar technologies for commercial purposes, are reserved. ISSN 1824-8039 . Published by SISSA Medialab.

PoS(ISGC2026)026

1. Introduction

Football data provide a demanding benchmark for evaluating Machine Learning (ML) models. Common learning tasks include goal-scoring prediction, action valuation, possession-value estimation, event forecasting, and tactical-state recognition. Besides the relevance in terms of sports analytics application, these tasks present several well-known ML challenges like severe class imbalance and rare positive events, temporally correlated samples, delayed and noisy labels, high-dimensional spatio-temporal covariates, multimodal event and positional representations, and strong feature interactions between action type, field location, game state, player quality, and match context. For this reason, football data represent a compelling benchmark also from an ML methodological research, and can be seen as an empirical laboratory for stress-testing robust supervised learning, probability estimation, and transfer-learning methods.

A central difficulty is that predictive models in football analytics are often brittle when deployed out of distribution. A model fitted on a source set of competitions may lose discrimination or calibration when applied to different leagues or seasons. This is an intrinsic issue, as competitions differ in playing style, tempo, tactical conventions, squad quality, defensive pressure, officiating patterns, and season-specific tactical trends. In statistical terms, the joint distribution of covariates and labels changes between source and target domains. This makes football a natural setting for transfer learning and domain adaptation, where the question is how to reuse information learned from a label-rich source domain in a related but shifted target domain [7–9].

The same setting exposes the three standard forms of dataset shift. Let X denote the feature representation of a game state and Y the binary short-horizon scoring label. *Prior shift* occurs when $P_t(Y) \neq P_s(Y)$: in our task, the base rate of scoring within the next actions may vary across competitions, seasons, and tournament formats. *Covariate shift* occurs when $P_t(X) \neq P_s(X)$: leagues and seasons can differ in action frequencies, field progression, spatial occupation, directness, and pressing-induced turnover patterns. *Concept shift* occurs when $P_t(Y \mid X) \neq P_s(Y \mid X)$: the same local action sequence may have different scoring implications under different defensive organization, transition dynamics, or finishing quality. In realistic football data these shifts are not mutually exclusive. They co-occur, which makes the domain harder but also more useful as a benchmark than settings in which a single shift type is isolated by construction.

We instantiate this benchmark using publicly available StatsBomb Open Data processed through the SPADL representation and the VAEP action-valuation formulation [1–4]. Each sample is a three-action game state, formed by the current action and the two immediately preceding ones, and the prediction target is whether the team in possession will score within the next ten actions. We use XGBoost as a strong tabular baseline and study a practical warm-start fine-tuning protocol rather than introducing a new architecture [11]. The purpose is diagnostic: to quantify when source knowledge transfers, how much target supervision is needed, and which aspects of performance improve under adaptation. Our contributions are threefold:

- we frame football data as a methodological benchmark for transfer learning under simultaneous prior, covariate, and concept shift, rather than only as a domain-specific prediction problem;
- we define a reproducible source–target protocol that separates time shift, league shift, and combined time–league shift using StatsBomb data and a SPADL/VAEP action-based pipeline;

PoS(ISGC2026)026

- we provide preliminary empirical evidence that warm-start fine-tuning of gradient-boosted trees yields mild but consistent gains in ranking metrics under limited target-label budgets.

2. Related work

2.1 Football analytics, action representations, and goal-scoring prediction

Football analytics has increasingly moved from aggregate match statistics and isolated outcome events toward action-based, context-aware modelling. Early and widely used indicators such as goals, assists, shots, or expected goals provide useful summaries, but they capture only a small fraction of the decisions that shape a match. Recent approaches aim to evaluate complete sequences of on-ball actions and estimate how the current game state affects future scoring opportunities.

StatsBomb Open Data provides access to event-level football data across several competitions, while the `socceraction` ecosystem offers tools for converting raw event streams into standardized action representations and for implementing action-valuation models [4]. A central component of this ecosystem is SPADL, the Soccer Player Action Description Language [4] that represents a match as a sequence of on-ball player actions with fixed attributes such as time, location, player, team, action type, body part, and result, thereby enabling ML models to operate on consistent feature representations. Within this representation, VAEP, Valuing Actions by Estimating Probabilities, assigns value to each action by measuring how it changes the probability that the team in possession will score or concede in the near future. Action valuation is therefore cast as a supervised problem over future scoring and conceding outcomes.

Several related works extend this idea of context-aware action valuation. Deep reinforcement-learning approaches have been used to learn action-value functions from soccer play-by-play data and to derive player evaluation metrics by aggregating the estimated impact of individual actions [5]. More recent football-AI systems exploit richer spatio-temporal structure, using graph and geometric deep learning to model tactical situations, predict likely outcomes, and recommend adjustments [6]. These works show that modern football analytics increasingly treats prediction and valuation as state-dependent problems, where the aim is not only to model goals but to estimate the future consequences of actions from game states.

The present work follows this action-based perspective, but focuses on a different methodological question. We use public StatsBomb data processed through a SPADL/VAEP pipeline and study whether a model trained to predict short-horizon scoring outcomes transfers across competitions, seasons, and leagues. In this sense, football action data provide both the task representation and the empirical setting for studying cross-domain generalization.

2.2 Transfer learning, dataset shift, and domain adaptation

Transfer learning studies how knowledge acquired in one learning problem can be reused to improve performance in another related problem [7]. This is especially useful when the source domain contains abundant labelled data, while the target domain is smaller, noisier, or more expensive to annotate. The key assumption is that the two domains are not identical but are related enough that information learned from the source remains useful in the target. Depending on the setting, transfer may involve changes in the input distribution, the task, or the label space, or a combination of these factors.

PoS(ISGC2026)026

Domain adaptation is a form of transfer learning in which the source and target domains share the same prediction task but differ in their data-generating distributions. In this case, the goal is to preserve or improve predictive performance when a model trained under one distribution is deployed under another. This situation is common in practical ML: training data are often collected in one environment, while deployment takes place in a related but shifted environment. Domain adaptation methods address this mismatch by using source data together with some form of target-domain information, which may be unlabelled target samples, a small labelled target set, or both. The need for domain adaptation arises from dataset shift. A source domain and a target domain may differ in the marginal distribution of labels, in the marginal distribution of inputs, or in the conditional relation between inputs and labels. These cases are commonly referred to as prior shift, covariate shift, and concept shift, respectively [8, 9]. In practice, these forms of shift are often not isolated. They may occur simultaneously, making transfer difficult because the model must cope not only with different input patterns, but also with different base rates and decision rules.

Several families of methods have been proposed to address these issues [10]. The present work follows a supervised adaptation strategy: we use a source-trained XGBoost model [11] and study whether warm-start fine-tuning on a limited labelled target set improves over both zero-shot transfer and target-only training from scratch.

3. Experimental design

The goal of the experimental design is to evaluate whether action-based prediction models generalize under controlled domain-shift scenarios. We consider a supervised goal-scoring prediction task built from SPADL/VAEP action sequences and use XGBoost as a strong tabular baseline. The study is designed to isolate three practically relevant forms of distribution shift: temporal shift across seasons, competition shift across leagues, and their combination. We then compare zero-shot transfer, target-only learning, and warm-start fine-tuning under limited target-label budgets.

Our experiments rely on publicly available football event data distributed by StatsBomb and processed through the `socceraction` stack. The current repository snapshot uses a SPADL-based action representation and a curated set of 138 VAEP features extracted from action sequences. The preprocessing pipeline is the following:

1. raw StatsBomb event streams are converted into SPADL actions;
2. actions are re-oriented so that the team in possession attacks in a consistent left-to-right direction;
3. VAEP game states are built from the last three actions;
4. feature functions are applied to obtain action-type, result, body-part, location, movement, game-score, and timing covariates;
5. the resulting features are merged back with competition metadata and labels.

3.1 Data and domain shift

The empirical analysis is based on a curated subset of StatsBomb Open Data [1]. The public repository does not provide uniform coverage across competitions and seasons: some domestic

PoS(ISGC2026)026

Competition	Seasons	Games
Premier League	2003–2015	418
La Liga	1973–2020	868
Serie A	1986–2015	381
Bundesliga	2015–2023	340
Ligue 1	2015–2022	435
Champions League	1970–2018	18
Europa League	1988–1988	3

Table 1: Coverage of the StatsBomb Open Data subset used in this study. The table reports, for each competition, the range of available seasons and the total number of games.

leagues are available over many seasons, whereas other competitions contain only a small number of matches. Table 1 summarizes the data used in the study.

For the exploratory data analysis in this section, we consider only competitions with at least one season containing more than ten available games. This provides enough season-level support for meaningful descriptive comparisons across leagues. The predictive modelling experiments, instead, use all available competitions listed in Table 1. In this case, sparse competitions are retained because they still contribute target-domain examples for evaluating transfer under realistic data availability.

Prior shift. A first source of domain shift is visible at the level of the marginal scoring distribution. Figure 1 compares the distribution of goals per match across the top-five European leagues in the 2015/2016 season. Although the task definition is identical across competitions, the base scoring profile is not. The average number of goals per match ranges from 2.52 in Ligue 1 and 2.58 in Serie A to 2.83 in the Bundesliga, with the Premier League (2.70) and La Liga (2.74) in between. The quartiles further show that the leagues differ not only in their average scoring rate, but also in the shape of their scoring distributions. Serie A and Ligue 1 have $q_{25} = 1$ and median $q_{50} = 2$, meaning that at least half of their matches contain at most two goals. By contrast, La Liga and the Bundesliga have $q_{25} = 2$ and median $q_{50} = 3$, indicating a distribution shifted toward higher-scoring matches. The Premier League has median $q_{50} = 3$, but $q_{25} = 1$, suggesting a more heterogeneous distribution with both low-scoring and high-scoring games. The upper quartile is $q_{75} = 4$ for Serie A, the Premier League, La Liga, and the Bundesliga, but only $q_{75} = 3$ for Ligue 1, showing that high-scoring matches are less frequent in Ligue 1 than in the other leagues.

Prior shift can also emerge over time within the same competition. Figure 2 shows the temporal evolution of average goals per game in La Liga from 2010/2011 to 2020/2021. The season-level averages vary substantially across years, ranging from lower-scoring seasons such as 2015/2016 to higher-scoring seasons such as 2012/2013 and 2016/2017. The fitted local trends further highlight changes in the scoring prior across predefined temporal windows. In the first window, from 2010/2011 to 2012/2013, the fitted average increases by $\Delta = +1.10$ goals per game, indicating a clear upward shift in the scoring environment. In the second window, from 2014/2015 to 2016/2017, the fitted change is also positive, $\Delta = +0.55$. In the last window, from 2018/2019 to 2020/2021, the fitted change is small, $\Delta = +0.05$, suggesting a comparatively stable scoring prior.

PoS(ISGC2026)026

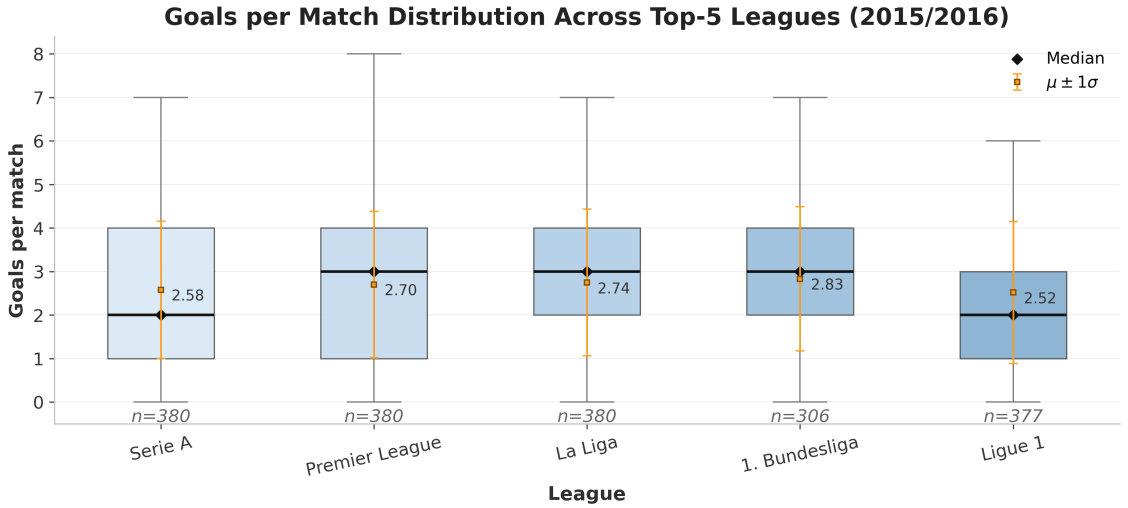


Figure 1: Evidence of prior shift across the top-five European leagues in the 2015/2016 season. The boxplots show the distribution of goals per match for each league, while the markers report quartiles, medians, and mean ± one standard deviation.

These differences indicate that some target domains are intrinsically more goal-prone than others, both across competitions and across seasons of the same competition.

Covariate shift. To assess covariate shift, we characterize league-specific playing styles through a clustering-based analysis of passing actions. We consider all actions labelled as `pass` or `cross` and represent each action using a compact set of spatial and directional features: the starting location of the pass, its length, and two normalized directional components. The starting coordinates indicate where the action originates on the pitch; the length captures how far the ball is moved; and the directional components encode the orientation of the pass independently of its length, distinguishing, for example, vertical, horizontal, and diagonal progressions. After standardizing these features, we apply K-Means to the pooled set of passes and crosses from all leagues, so that clusters correspond to common passing patterns shared across the whole dataset. We then compute, for each league, the frequency with which each cluster appears relative to that league's total number of passes. These cluster proportions are transformed into league-level z-scores with respect to the overall cluster distribution. A high positive z-score indicates that a given passing pattern is used substantially more often in that league than in the pooled reference distribution. For visualization, we select the five clusters with the largest positive z-scores for each league, i.e., the clusters that are most distinctive of that league. Within each selected cluster, we plot 30 representative passes close to the corresponding cluster centroid to summarize the typical spatial realization of that pattern. Figure 3 shows that the distribution of passing covariates differs substantially across competitions.

La Liga displays a relatively high spatial barycenter, with characteristic clusters located closer to the opponent's half and with several short or medium-range passing patterns. This suggests a style based on shorter combinations, advanced possession, and a higher team block, which is consistent with stronger pressure on the opponent's defensive structure. By contrast, the Bundesliga shows more distinctive long-pass patterns, including vertical and diagonal progressions that cover larger

PoS(ISGC2026)026

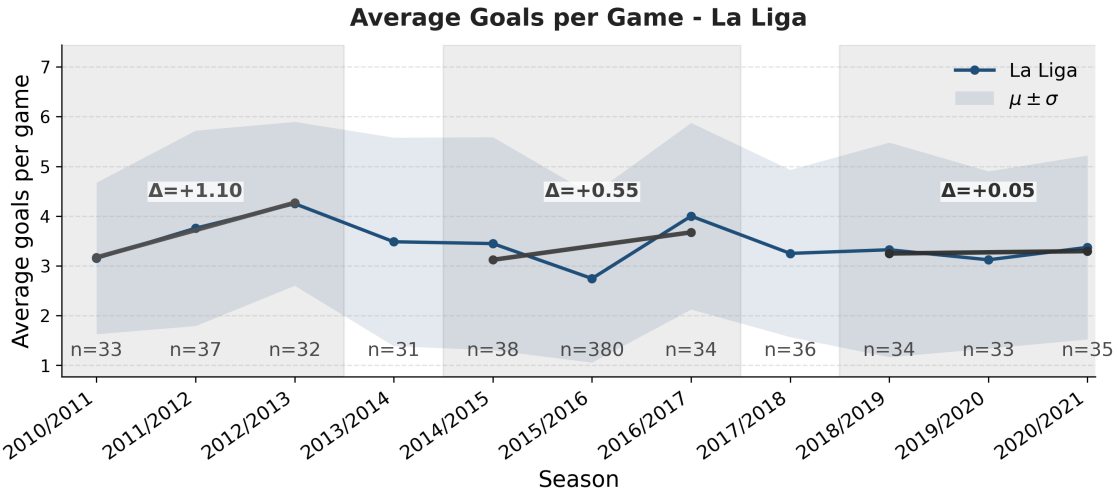


Figure 2: Temporal evidence of prior shift in La Liga. The figure shows the season-level average goals per game from 2010/2011 to 2020/2021. The blue line with circular markers reports the mean goals per game for each season, while the light-blue ribbon represents inter-match variability within each season, computed as mean ± one standard deviation. Gray vertical bands indicate three predefined comparison windows. Within each window, the dark segment shows a least-squares linear trend fitted to the seasonal means, and the annotation reports the corresponding start-to-end change Δ. The n values at the bottom indicate the number of matches available for each season.

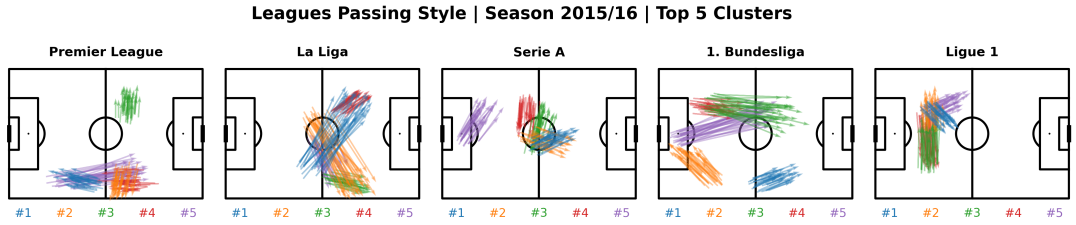


Figure 3: Evidence of covariate shift through league-specific passing-style clusters. Passes are represented by their start and end coordinates, standardized, and clustered with K-Means to identify recurrent spatial passing patterns. For each league, cluster frequencies are converted into league-level z-scores, and the five most distinctive clusters are visualized. The differences between leagues indicate that the covariate distribution of action-based game states varies across competitions, providing empirical evidence of covariate shift.

distances and connect deeper or central areas with advanced zones. The Premier League presents characteristic wide and deeper passing patterns together with some central forward progression, suggesting a more heterogeneous passing structure. Serie A shows more compact clusters around midfield and right-sided central areas, indicating a more localized and structured circulation of the ball. Ligue 1 presents several distinctive patterns concentrated in deeper and middle-third regions, suggesting a more cautious build-up structure compared to others.

A model trained on one league may therefore be evaluated on target data whose passing and progression patterns are systematically different from those observed during training, motivating the need for cross-league and year–league transfer experiments.

PoS(ISGC2026)026

Concept shift. Finally, we analyse concept shift by comparing the action patterns that lead to positive scoring outcomes across leagues. Figure 4 reports the most frequent three-action sequences ending in a goal-related shot event, grouped by competition. Each triplet describes the last three on-ball actions before the scoring event, such as `pass -> dribble -> shot` or `dribble -> cross -> shot`. Unlike the previous covariate-shift analysis, which focuses on how the distribution of input actions changes across leagues, this analysis focuses on how different action patterns are associated with successful scoring outcomes.

The most common goal-ending pattern, `pass -> dribble -> shot`, is frequent in all leagues, accounting for roughly 18%–19% of goal-ending triplets. However, the relative importance of other patterns differs substantially across competitions. For example, `dribble -> cross -> shot` is more prominent in La Liga and the Bundesliga than in Serie A or the Premier League, suggesting that wide or crossing-based attacking sequences contribute differently across leagues. Similarly, `dribble -> pass -> shot` is frequent in La Liga and Serie A but less frequent in Ligue 1, indicating that central combination play before shots has different relevance depending on the competition. Penalty-related sequences, such as `dribble -> foul -> shot_penalty`, are more visible in Serie A than in several other leagues, while `take_on -> dribble -> shot` is particularly pronounced in Ligue 1 and the Bundesliga. These differences suggest that the same scoring label can arise from different tactical mechanisms across domains.

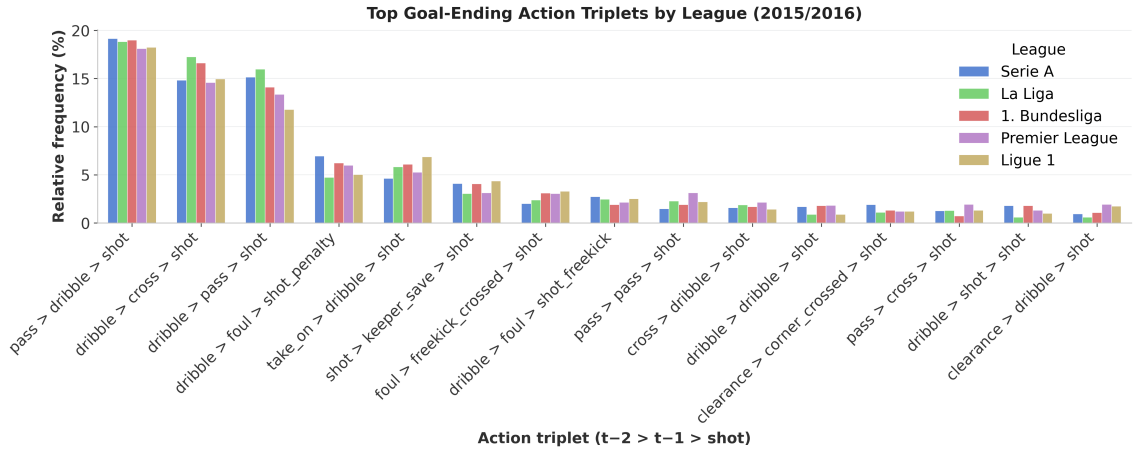


Figure 4: Evidence of concept shift through goal-ending action triplets. The figure reports the relative frequency of the most common three-action sequences preceding goal-related shot events in the 2015/2016 season, separately for each league. These differences indicate that the action patterns associated with positive scoring outcomes vary across competitions, providing empirical evidence that the conditional relationship between game-state covariates and scoring labels may change across domains.

3.2 Prediction target and domain-shift benchmark

We focus on the `scores` label used in VAEP-style action valuation. Denote by

$$s_t = (a_{t-2}, a_{t-1}, a_t)$$

the game state formed by the current action and the two immediately preceding actions. Let

$$y_t = \mathbb{1}\{\text{team in possession after } a_t \text{ scores within the next } k \text{ actions}\},$$

PoS(ISGC2026)026

with $k = 10$, following the standard VAEP formulation. The learning problem is therefore

$$p_\theta(y_t = 1 \mid s_t) = f_\theta(x_t),$$

where x_t is the tabular feature representation of the game state s_t. In the present study, we report results for the `scores` target, although the same pipeline can also be applied to the complementary `concedes` label. This prediction task is particularly suitable for studying domain shift. First, the positive class is rare, since only a small fraction of game states are followed by a goal within the next ten actions. This makes the problem a challenging rare-event prediction task and prevents trivial evaluation based on raw accuracy. Second, the same target definition can be applied consistently across leagues and seasons, which allows us to compare source and target domains under a common supervised-learning formulation. Third, the task is directly affected by the three forms of dataset shift discussed above.

Prior shift arises because the base rate of short-horizon scoring can vary across leagues, seasons, and tournament formats. Covariate shift arises because the distribution of action-based features changes across competitions: teams differ in action frequencies, field progression patterns, passing styles, and spatial occupation. Concept shift may also arise because the same local configuration of actions can imply different scoring probabilities depending on the tactical and defensive context of a given competition. For example, a three-action sequence that is dangerous in one league may be less predictive of scoring in another if defensive pressure, tempo, or attacking conventions differ.

Thus, the task combines a fixed prediction target with naturally changing data-generating distributions. This makes short-horizon goal-scoring prediction a useful benchmark for transfer learning and domain adaptation: a model trained on a label-rich source competition must be tested on target domains where the scoring prior, the input distribution, and the relation between game states and scoring outcomes may all change simultaneously.

3.3 Transfer-learning protocol

We define a source–target protocol to evaluate whether knowledge learned from a label-rich football domain can be transferred to related but shifted target domains. The source domain $\mathcal{D}_s$ is men's La Liga, using all available seasons except 2015/16, 2019/20, and 2020/21. These excluded seasons are kept out of source training and used as part of the target evaluation. The target domain $\mathcal{D}_t$ consists of major European men's competitions, namely Premier League, Bundesliga, Serie A, UEFA Champions League, UEFA Europa League, and the held-out La Liga seasons.

This design induces three evaluation regimes. The *time-shift* regime evaluates transfer from source La Liga seasons to held-out La Liga seasons, thus isolating temporal drift within the same competition. The *league-shift* regime evaluates transfer to other competitions in seasons that overlap with the temporal support of the source data, mainly the 2015/16 season and additional competition-specific overlaps. The *time–league-shift* regime evaluates transfer to other competitions in non-overlapping seasons, combining competition shift with temporal drift. In all cases, the target data are split into a labelled adaptation subset and a held-out test subset.

We compare three training strategies. In *source-only zero-shot transfer*, an XGBoost model is trained on the source domain and then evaluated directly on the target test set, without using any labelled target examples. This setting measures how well the source model generalizes out

PoS(ISGC2026)026

of domain. In *target-only training from scratch*, a new XGBoost model is trained only on the labelled target adaptation subset and evaluated on the corresponding target test set. This baseline measures what can be learned from target labels alone, without any source-domain pretraining. In *warm-start fine-tuning*, the source-trained XGBoost model is used as initialization and further trained on the labelled target adaptation subset before evaluation. This setting tests whether limited target supervision can adapt the source decision function to the target domain.

A central component of the protocol is the *target-label budget*. Since the target domain is assumed to be label-scarce, we do not always allow the model to use all target games for adaptation. Instead, for each target regime we sample only a fixed fraction of target games as labelled adaptation data. We consider budgets of 10%, 20%, and 30% of the available target games. The remaining target games are kept as a held-out test set and are never used during adaptation. Importantly, budgets are defined at the game level rather than at the action level, so that actions from the same match cannot appear simultaneously in adaptation and test data.

For the target-only and fine-tuning strategies, the sampled adaptation subset is further split into training and validation partitions. The training partition is used to fit the target-only model or to continue training the source model, while the validation partition is used for early stopping and threshold selection. The source-only zero-shot baseline does not use the target adaptation labels; it is evaluated directly on the same target test split for comparability.

To quantify the variability induced by limited target supervision, the entire target sampling procedure is repeated five times with different random seeds for each label budget and evaluation regime. We therefore report performance as mean and standard deviation across seeds. This makes it possible to distinguish systematic transfer effects from fluctuations due to the particular choice of labelled target games.

3.4 Metrics

Because the positive class is rare, we treat PR-AUC as the primary ranking metric and use ROC-AUC as a complementary discrimination measure. In particular, we report PR-AUC, ROC-AUC, Brier score, and F1, and we monitor precision, recall, and log-loss as secondary diagnostics [12–15]. Given the extreme class imbalance, looking at combined metrics is particularly important to get a comprehensive view of model performance. For thresholded classification, we select the decision threshold on the validation split by maximizing F1 over a bounded threshold grid.

4. Results

4.1 Source-domain validation

Before analysing transfer across domains, we first evaluate the difficulty of the source-domain prediction task. Table 2 reports validation performance on the source domain in terms of positive prevalence, precision, recall, F1, ROC-AUC, and PR-AUC. The positive class accounts for only about 1.2% of the samples, confirming that the task is strongly imbalanced. This makes raw accuracy uninformative and motivates the use of ranking metrics, threshold-based metrics, and probabilistic metrics.

Even in-domain, the task remains non-trivial. The model achieves a ROC-AUC of 0.89, showing that it learns meaningful ranking information, but the PR-AUC is substantially lower

PoS(ISGC2026)026

Split	Positives (%)	Precision	Recall	F1	ROC-AUC	PR-AUC
Source validation	~ 1.2	0.3780	0.2568	0.3059	0.8917	0.2833

Table 2: Validation performance on the source domain reported in the conference presentation. Even in-domain, the task is highly imbalanced and non-trivial.

(0.28), as expected under severe class imbalance. Precision, recall, and F1 further show that identifying positive scoring events is difficult despite the use of a strong tabular model. These results indicate that the downstream transfer-learning experiments are performed on an intrinsically challenging rare-event prediction task, rather than on an artificially easy source-domain problem.

4.2 Cross-domain transfer performance

We evaluate whether source-domain knowledge transfers to shifted target domains. The comparison involves three strategies: source-only zero-shot transfer, target-only training from scratch, and warm-start fine-tuning. Figure 5 summarizes the results across different target-label budgets for four evaluation settings: season, league, combined league–season, and global target shifts.

The global target-shift setting provides the clearest evidence in favour of fine-tuning. Across target-label budgets, warm-start fine-tuning is consistently competitive with, and often slightly better than, the source-only zero-shot baseline. The absolute gains are small, but they are systematic in the ranking and classification metrics. This suggests that the source model captures useful general football structure, while a limited amount of target supervision can still adjust the decision function to the target distribution. The individual shift settings show a more variable picture. In the season-shift, league-shift, and league–season-shift regimes, fine-tuning does not always clearly dominate the source-only model. A plausible explanation is that these splits contain fewer target games, so the adaptation subset may provide limited additional information beyond the source-domain pretraining. As a consequence, the effect of fine-tuning is more sensitive to the particular target games sampled under each random seed.

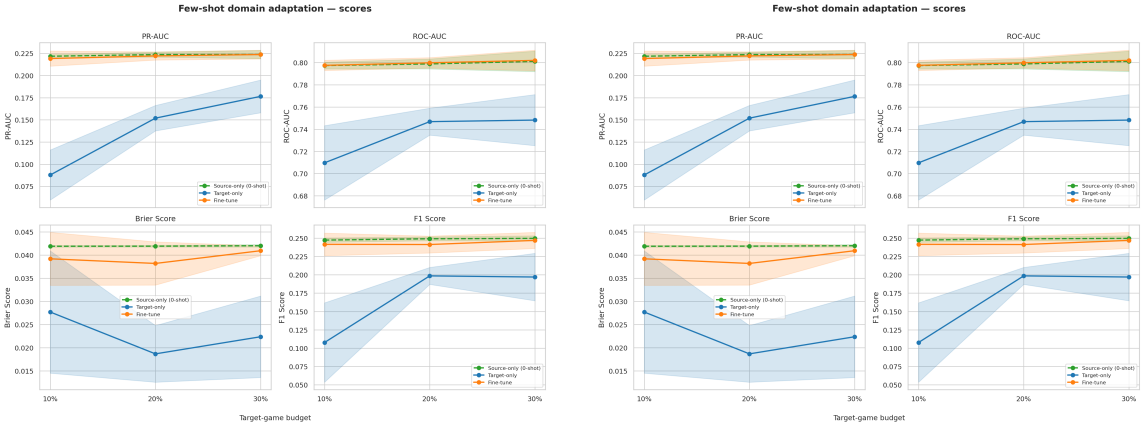
The target-only baseline follows the expected trend. At small label budgets, training from scratch on target data is generally weaker than using the source-trained model, because the model does not benefit from source-domain information. As the target-label budget increases, target-only training progressively closes the gap. In some settings, it becomes competitive with the transfer-based approaches. This confirms that transfer is most valuable in the low-label regime, whereas direct target training becomes more viable as more labelled target games are available.

Across metrics, the improvements are clearer for PR-AUC, ROC-AUC, and F1. These metrics capture positive-class ranking, global ranking discrimination, and threshold-based classification performance. By contrast, the Brier score does not show the same systematic improvement under fine-tuning. This indicates that adaptation can improve classification behaviour, but does not automatically produce better calibrated probabilities under domain shift.

4.3 Global target-shift summary

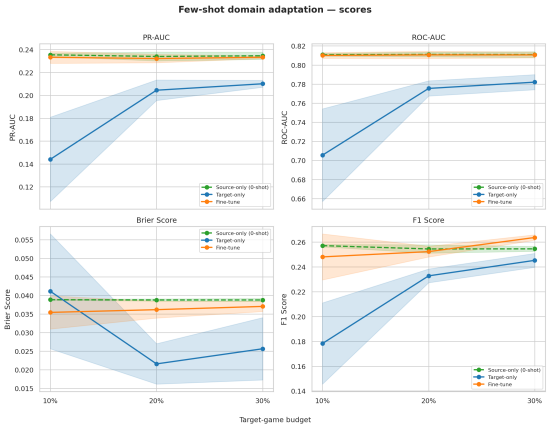
To complement the visual comparison, Table 3 reports the numerical results for the global target-shift setting. Values are reported as mean and standard deviation across the repeated target-

PoS(ISGC2026)026

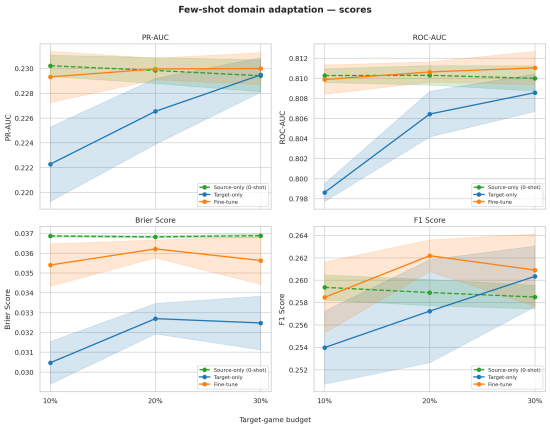


(a) Season shift

(b) League shift



(c) League–season shift



(d) Global target shift

Figure 5: Few-shot domain-adaptation results for the scores task. Each panel reports performance as a function of the target-game budget for source-only zero-shot transfer, target-only training from scratch, and warm-start fine-tuning. The four panels correspond to season shift, league shift, combined league–season shift, and global target shift. Shaded regions indicate variability across random target-game samples (seed of the splitting).

game samples. The table confirms the trends observed in Figure 5. Fine-tuning achieves the best or near-best PR-AUC and ROC-AUC across budgets, while target-only training improves steadily as the target-label budget increases. At the same time, the Brier score and log-loss are generally lower for the target-only model, showing that better discrimination and better calibration do not necessarily coincide.

5. Conclusion

We presented a proof-of-concept study of transfer learning for football analytics under realistic cross-domain shift. Using public StatsBomb event data processed through a SPADL/VAEP action-based pipeline, we formulated short-horizon goal-scoring prediction as a supervised rare-event task and evaluated whether a model pretrained on men's La Liga can transfer to shifted target domains

PoS(ISGC2026)026

Budget	Strategy	PR-AUC	ROC-AUC	Brier
10%	Source-only (0-shot)	**0.2302 ± 0.0009**	**0.8103 ± 0.0007**	0.0369 ± 0.0001
10%	Target-only	0.2223 ± 0.0030	0.7986 ± 0.0009	**0.0305 ± 0.0011**
10%	Fine-tune	0.2293 ± 0.0021	0.8099 ± 0.0015	0.0354 ± 0.0011
20%	Source-only (0-shot)	0.2298 ± 0.0010	0.8103 ± 0.0010	0.0368 ± 0.0001
20%	Target-only	0.2265 ± 0.0027	0.8064 ± 0.0023	**0.0327 ± 0.0008**
20%	Fine-tune	**0.2299 ± 0.0009**	**0.8106 ± 0.0010**	0.0362 ± 0.0005
30%	Source-only (0-shot)	0.2294 ± 0.0013	0.8100 ± 0.0013	0.0369 ± 0.0001
30%	Target-only	0.2295 ± 0.0014	0.8086 ± 0.0019	**0.0325 ± 0.0014**
30%	Fine-tune	**0.2300 ± 0.0013**	**0.8110 ± 0.0017**	0.0356 ± 0.0012

Table 3: Held-out global target-shift performance for source-only zero-shot transfer, target-only learning, and warm-start fine-tuning. Values are mean ± standard deviation over repeated target-game samples. Bold values indicate the best score within each budget and metric.

covering held-out La Liga seasons and other major European men's competitions. The protocol separates temporal shift, league shift, and combined league–season shift, and compares three practically relevant strategies: source-only zero-shot transfer, target-only training from scratch, and warm-start fine-tuning under limited target-label budgets.

The results support two main conclusions. First, football action data provide a useful methodological benchmark for domain adaptation. The task is difficult even in-domain because positive scoring events are rare, labels are noisy, and performance must be assessed beyond raw accuracy. At the same time, the target domains differ from the source in meaningful and interpretable ways, including scoring prevalence, action distributions, and competition-specific scoring mechanisms. This makes the source–target mismatch realistic rather than artificial. Second, warm-start fine-tuning provides modest but consistent benefits in the low-label regime, especially for ranking and threshold-based metrics such as PR-AUC, ROC-AUC, and F1. The gains are clearest in the global target-shift setting, while the individual shift regimes show more variability, likely because some target splits contain relatively few games and therefore provide limited additional information for adaptation. Target-only models follow the expected behaviour: they are weaker when labelled target data are scarce, but progressively close the gap as the target-label budget increases. This suggests that transfer is most useful when target supervision is limited, whereas direct target training becomes more competitive as more labelled target games are available.

The study also highlights important limitations. The evidence is preliminary, relies on a single model family, and considers one adaptation mechanism: continuing XGBoost training from a source-trained booster. Moreover, the current feature representation is intentionally action-based and does not yet exploit richer off-ball or 360 positional context. Finally, calibration remains an open issue. Fine-tuning can improve discrimination without necessarily improving probability quality, as shown by the less systematic behaviour of Brier score and log-loss.

Future work should therefore extend the benchmark in three directions. First, domain shift should be measured more explicitly, for example through domain classifiers, feature-discrepancy measures, temporal-drift analyses, and ablations that isolate which components of the football domain drive transfer failure. Second, richer representations should be explored, including 360-

PoS(ISGC2026)026

derived features, sequence models, and graph-based models that better capture spatial and tactical context. Third, the adaptation methods should be broadened beyond warm-start boosting to include importance weighting, multi-domain training, domain-invariant representations, meta-learning over competitions, and dedicated post-hoc or shift-aware calibration methods.

Overall, this work argues that football is not only an application domain for machine learning, but also a realistic and interpretable benchmark for transfer learning and domain adaptation. It combines severe class imbalance, structured action semantics, temporal drift, and competition heterogeneity in a public and reproducible setting. This makes it a valuable testbed for studying how predictive models behave when deployed outside the distribution on which they were trained.

Code availability. The implementation accompanying this work is publicly available at: https://github.com/clissa/transfer-learning-for-football-AI

Funding. This research was partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI" and the European Commission under the NextGeneration EU programme. The collection of original images was supported by funding from the University of Bologna and the European Space Agency (Research agreement collaboration 4000123556).

References

[1] **StatsBomb**, *StatsBomb Open Data*, GitHub repository, 2019. https://github.com/statsbomb/open-data.

[2] **socceraction developers**, *socceraction: A Soccer Action Valuation Toolkit*, documentation, version 1.5.3. https://socceraction.readthedocs.io/en/latest/.

[3] **socceraction developers**, *VAEP documentation*, https://socceraction.readthedocs.io/en/latest/documentation/valuing_actions/vaep.html.

[4] T. Decroos, L. Bransen, J. Van Haaren and J. Davis, *Actions Speak Louder than Goals: Valuing Player Actions in Soccer*, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1851–1861, doi: 10.1145/3292500.3330758.

[5] G. Liu, Y. Luo, O. Schulte and T. Kharrat, *Deep soccer analytics: learning an action-value function for evaluating soccer players*, *Data Mining and Knowledge Discovery* **34** (2020) 1531–1559, doi: 10.1007/s10618-020-00705-9.

[6] Z. Wang et al., *TacticAI: an AI assistant for football tactics*, *Nature Communications* **15** (2024) 1906, doi: 10.1038/s41467-024-45965-x.

[7] S.J. Pan and Q. Yang, *A Survey on Transfer Learning*, *IEEE Transactions on Knowledge and Data Engineering* **22** (2010) 1345–1359, doi: 10.1109/TKDE.2009.191.

PoS(ISGC2026)026

[8] J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla and F. Herrera, *A unifying view on dataset shift in classification*, *Pattern Recognition* **45** (2012) 521–530, doi: 10.1016/j.patcog.2011.06.019.

[9] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer and N.D. Lawrence eds., *Dataset Shift in Machine Learning*, MIT Press, Cambridge MA, 2009.

[10] M. Sugiyama, M. Krauledat and K.-R. Müller, *Covariate Shift Adaptation by Importance Weighted Cross Validation*, *Journal of Machine Learning Research* **8** (2007) 985–1005.

[11] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[12] T. Saito and M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*, *PLOS ONE* **10** (2015) e0118432, doi: 10.1371/journal.pone.0118432.

[13] J. Platt, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*, in *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.

[14] A. Niculescu-Mizil and R. Caruana, *Predicting Good Probabilities with Supervised Learning*, in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[15] G.W. Brier, *Verification of Forecasts Expressed in Terms of Probability*, *Monthly Weather Review* **78** (1950) 1–3, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[16] L. Clissa and A. Macaluso, *transfer-learning-for-football-AI*, GitHub repository, 2026. https://github.com/clissa/transfer-learning-for-football-AI.

PoS(ISGC2026)026