Applied Data Science and Analysis
Vol.2026, **pp**. 91–110
DOI: https://doi.org/10.58496/ADSA/2026/006; ISSN: 3005-317X
https://mesopotamian.press/journals/index.php/ADSA

Research Article

A Big Data Analytics Framework for Decision-Making in Sports Performance Optimization

Bharti Gawali [1,*], Mohammed Basil Abdulkareem [2], Munther H. Abed [3]

[1] *Department of computer science and information technology, University Chhatrapati Sambhaji Nagar, India.*

[2] *Department of Business Administration, College of Administration and economics, University of Anbar, Anbar, 31001, Iraq.*

[3] *Faculty of Computing, University Malaysia Pahang, Pahang, Malaysia.*

ARTICLE INFO

Article History
Received 08 Jun 2026
Revised 20 Jul 2026
Accepted 09 Aug 2026
Published 11 Sep 2026

Keywords
Sports analytics,
spatio-temporal event data,
big data frameworks,
expected goals,
Expected Threat,
temporal convolutional networks,
workload monitoring,
explainable machine learning.

ABSTRACT

Elite team sports now generate continuous, heterogeneous data streams from optical tracking, satellite-based wearables and manually annotated event feeds, yet most published analytics remain single-purpose models evaluated on proprietary corpora that cannot be reproduced. This paper proposes and empirically evaluates an end-to-end big data analytics framework that converts high-throughput spatio-temporal match data into decision support for coaching and performance staff. The framework comprises five layers: distributed ingestion, stream conditioning, a shared feature store, a multi-task learning engine, and an interpretable decision layer. It is instantiated on a fully public corpus of 1,517 matches and 5,321,459 events drawn from four European domestic seasons, and validated on three coupled tasks. First, an expected-goals module enriched with freeze-frame spatial descriptors reaches an AUC of 0.817 on 303 held-out matches, statistically indistinguishable from the commercial model shipped with the same data (0.819), while a distance-and-angle baseline reaches only 0.747. Second, a compact temporal convolutional network forecasts shot occurrence within a ten-second horizon at an AUC of 0.802 using 9,089 parameters, matching gradient boosting with a 50-fold smaller memory footprint and a median single-event latency of 0.034 ms. Third, a zone-based Expected Threat surface converges in 68 iterations and correlates with team-season goals at r = 0.859. An event-derived workload module shows that exponentially weighted load ratios are substantially more stable than conventional rolling ratios. The framework demonstrates that transparent, reproducible pipelines can match proprietary systems while remaining deployable at the edge.

1. INTRODUCTION

Performance analysis in association football was, for most of its history, a notational exercise. Analysts counted discrete outcomes — shots, passes, tackles — and aggregated them into match reports that discarded almost all contextual information about where and when an action occurred, who was nearby, and what alternatives were available to the player in possession [1]. The introduction of semi-automatic optical tracking and, subsequently, of satellite- and inertial-based wearable telemetry transformed the measurement problem. Modern systems sample the position of every player and the ball at rates between 10 and 25 Hz, producing on the order of one million spatial observations per match, while annotated event feeds add several thousand semantically labelled actions per fixture [2].

This shift creates a genuine big data problem rather than merely a larger statistics problem. The volume of a single professional season now runs to hundreds of gigabytes once tracking, event and physiological streams are combined. The velocity requirement is asymmetric: retrospective analysis may tolerate overnight batch processing, but in-match tactical support and live workload monitoring require decisions within the interval between consecutive frames. The variety dimension is equally demanding, because positional telemetry, discrete event annotations and athlete availability records differ in sampling rate, coordinate system, noise characteristics and semantic granularity. Satellite-based systems in particular introduce measurement error that varies with activity intensity and satellite geometry, and their validity has been the subject of extensive methodological scrutiny [3], [4].

Reviews of the field converge on a consistent diagnosis. Systematic surveys of tactical performance analysis report that the majority of published studies analyse a single season of a single club, use proprietary data that cannot be redistributed, and evaluate models without a held-out temporal or match-level partition [5]. Reviews of machine learning applications in professional football reach similar conclusions, noting that reported accuracy figures are rarely accompanied by

*Corresponding author. Email: drbhartirokade@gmail.com

computational cost, calibration analysis, or any assessment of deployability [6], [7]. The consequence is a literature rich in individual models but poor in integrated, reproducible systems.

The release of large open event corpora has begun to change this situation. A public collection of spatio-temporal match events covering seven competitions was released with full documentation in 2019 [8], and an openly licensed collection of annotated matches including shot freeze frames — snapshots of the position of every visible player at the instant of each shot — is maintained and periodically extended [9]. These resources make it possible, for the first time, to specify an analytics architecture and to verify every quantitative claim about it against data that any reader can download.

This paper takes that opportunity. Our contribution is fivefold. (i) We specify a five-layer big data architecture that unifies ingestion, stream conditioning, feature management, multi-task learning and interpretable decision support, and we state explicitly which layers are empirically evaluated and which are specified only. (ii) We instantiate the architecture on 1,517 matches and 5,321,459 events from four full European domestic seasons and report every result on match-level held-out partitions. (iii) We test generalisation directly, scoring the frozen models on 1,135 further matches from thirteen competitions that include tournament football and the women's game, none of which contributed to fitting. (iv) We evaluate three coupled decision-support tasks — shot valuation, real-time threat forecasting and workload monitoring — under a common protocol that reports predictive accuracy, calibration, inference latency and memory footprint together, because a model that cannot meet the latency budget of the touchline is not a decision-support model regardless of its area under the curve. (v) We benchmark our open shot-valuation model directly against the commercial model distributed with the same corpus, which provides an external reference point that is almost never available in this literature.

The remainder of the paper is organised as follows. Section 2 reviews the state of the art in action valuation, tactical modelling and workload management. Section 3 specifies the framework and its mathematical formulation. Section 4 reports the empirical evaluation. Section 5 discusses practical translation, interpretability and edge cases, and Section 6 concludes.

2. RELATED WORK

2.1 Valuing Actions and Possessions

The central methodological problem in football analytics is that goals are rare, so evaluating players and tactics by outcomes alone yields estimators with very high variance. The dominant response has been to assign every action a value equal to its effect on the probability of scoring. The VAEP framework values each on-ball action by the change it induces in the probabilities of scoring and conceding within a fixed horizon of subsequent actions, learned from a large corpus of labelled event sequences [10]. Expected Threat (xT) takes a complementary and more parsimonious route: the pitch is discretised into zones, and the value of a zone is defined self-consistently as the probability of eventually scoring from it, obtained as the fixed point of a Markov recursion over shooting, moving and turnover probabilities estimated directly from historical actions [11]. Because it requires only event data and has a closed interpretation, xT has become a standard component of possession-value systems.

Where full tracking data are available, richer formulations become possible. The expected possession value framework decomposes possession value into pass, ball-drive and shot components, each estimated by a dedicated deep architecture over the spatial configuration of all players, producing calibrated surfaces that can be inspected visually by practitioners [12]. This line of work extends the multiresolution stochastic process model originally developed for basketball possessions, which showed that possession value can be estimated coherently at every instant rather than only at discrete events [13]. For shot valuation specifically, models that synchronise event annotations with positional data and include defensive configuration and goalkeeper placement outperform purely geometric models by a clear margin [14], and rule-based formulations of instantaneous danger combining zone, control, pressure and defensive density have been validated on Bundesliga tracking data [15].

2.2 Workload Monitoring and Availability

The second research programme relevant to performance optimisation concerns the relationship between accumulated load and athlete availability. The acute-to-chronic workload ratio (ACWR) was popularised as a practical index of whether recent load is disproportionate to the load an athlete is conditioned for, and was associated with a "sweet spot" between roughly 0.8 and 1.3 [16]; associations between ACWR and injury risk have been reported in professional soccer [17]. Methodologically, however, the index has attracted sustained criticism. Exponentially weighted moving averages were proposed as a better-behaved alternative that respects the decaying influence of past sessions [18]. More fundamentally, the conventional coupled ratio shares its numerator with its denominator, which induces spurious correlation, and the ratio

formulation itself has poor statistical properties for causal interpretation [19]–[21]. Consensus guidance on load monitoring accordingly recommends reporting multiple representations of load rather than a single ratio [22].

Machine learning approaches to injury forecasting have been demonstrated on club-held GPS corpora, with interpretable tree ensembles achieving useful sensitivity on a single-season professional dataset [23]. Systematic review of this literature nevertheless finds small samples, heterogeneous outcome definitions and a near-total absence of external validation [24]. A structural obstacle underlies these findings: no openly licensed dataset pairs athlete workload with injury labels at the granularity the models require. We treat this obstacle explicitly rather than circumventing it, and Section IV-G reports what can and cannot be established from open data.

2.3 Positioning of the Present Work

Across both programmes, three gaps recur: single-task models rather than integrated pipelines; evaluation that ignores computational cost; and irreproducibility. The framework specified below addresses the first by sharing one feature store across three prediction tasks, the second by reporting latency and footprint alongside every accuracy figure, and the third by restricting the entire empirical study to a public corpus.

3. PROPOSED FRAMEWORK AND METHODOLOGY

3.1 Architectural Overview

The framework is organised into five layers with a single feedback path. Layer L1 ingests raw streams; L2 conditions them into a common spatio-temporal representation; L3 materialises features into a store shared by all downstream consumers; L4 hosts the learning engine; and L5 renders decisions for coaching and performance staff. Labelled outcomes and monitoring statistics flow back from L5 to L1 so that models can be refreshed as playing style and squad composition drift. Fig. 1 shows the architecture; the same structure is given below as a text flowchart for readers working from plain-text sources.

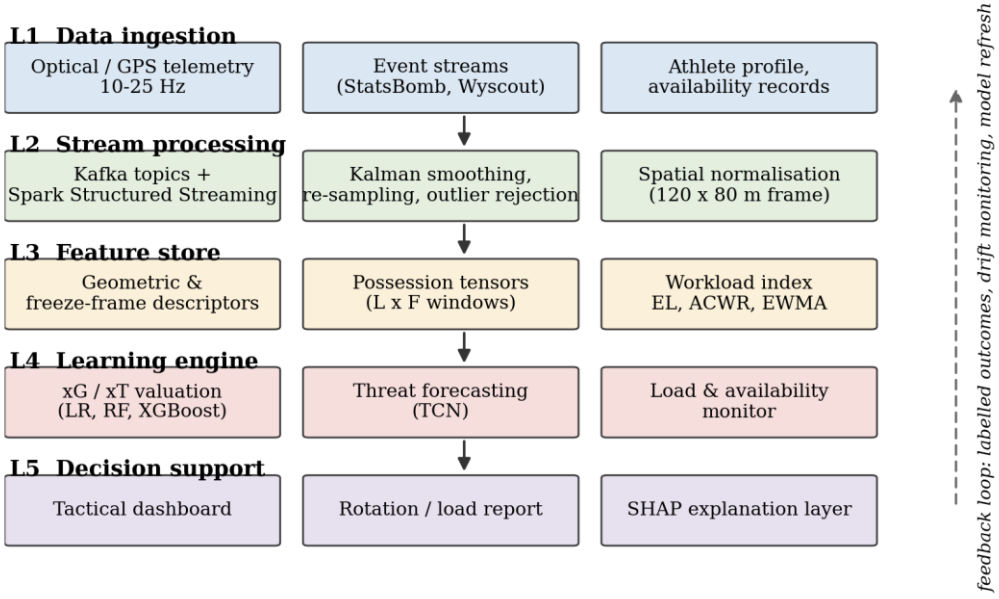

Fig. 1. Five-layer architecture of the proposed framework. Each column within a layer denotes a parallel processing path; the dashed arrow denotes the feedback loop used for model refresh.

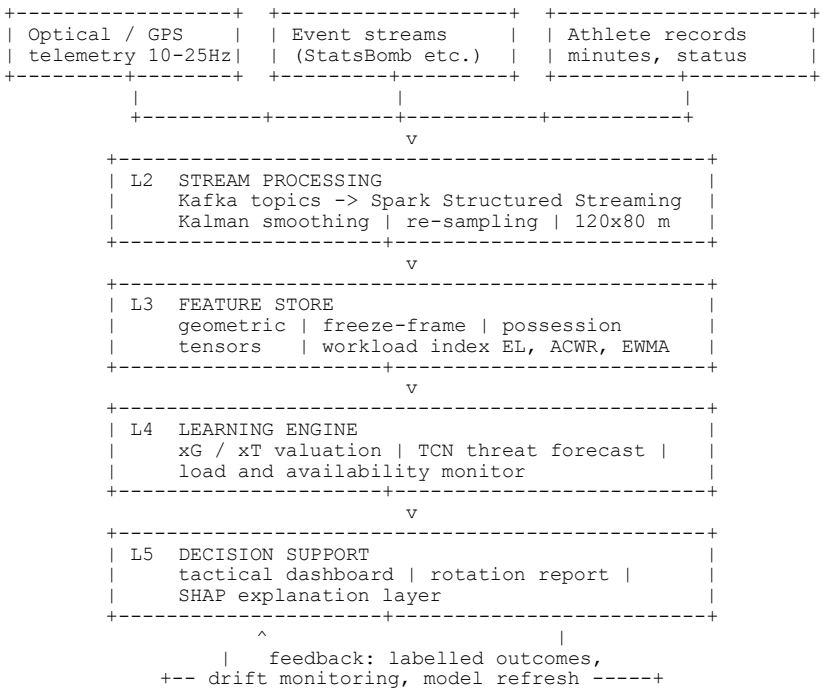

Fig. 2. Text rendering of the framework pipeline.

3.2 Data Ingestion and Corpus Description

The empirical instantiation uses the StatsBomb open data collection [9], which distributes fully annotated event records under an open licence. We selected the four competitions for which complete domestic seasons are available: the English Premier League, Spanish La Liga, Italian Serie A and French Ligue 1, all for 2015/16. Every match file was retrieved programmatically from the public repository and parsed in a streaming fashion without persisting raw payloads.

Each event carries a type from a controlled vocabulary of roughly forty categories, a pitch location on a 120 × 80 unit coordinate frame, a timestamp resolved to the second, the acting player and team, a possession identifier, and type-specific qualifiers. Shots additionally carry a freeze frame: the position and team membership of every player visible to the annotator at the instant of the shot, together with the vendor's own expected-goals estimate, which we retain solely as an external benchmark and never as a model input. Table I summarises the corpus. Penalty kicks were excluded from all shot models because their conversion probability is governed by a fixed geometry and carries no tactical information.

TABLE I. COMPOSITION OF THE EMPIRICAL CORPUS (STATSBOMB OPEN DATA, 2015/16 SEASON).

Competition	Matches	Events	Shots*	Goals*	Events per match
Premier League (England)	380	1,313,773	9,817	914	3,457
La Liga (Spain)	380	1,295,354	9,071	945	3,409
Serie A (Italy)	380	1,353,739	9,877	858	3,562
Ligue 1 (France)	377	1,358,593	8,723	852	3,604
Total	1,517	5,321,459	37,488	3,569	3,508

*Excluding penalty kicks. The overall conversion rate of non-penalty shots is 9.52 %. Raw event payloads occupy approximately 3.03 MB per match, or 4.6 GB for the full corpus.

3.3 Stream Conditioning

Layer L2 is specified around a publish-subscribe log and a distributed stream processor. Raw telemetry is published to partitioned topics keyed by match identifier, and consumed by micro-batch structured streaming jobs whose unified batch and streaming semantics allow the same transformation code to serve live and retrospective workloads [25], [26].

Partitioning by match preserves event ordering within a fixture while permitting horizontal scaling across concurrent fixtures.

Positional telemetry requires filtering before any kinematic quantity can be derived, because differentiating raw satellite positions amplifies measurement noise. We adopt a constant-velocity linear state-space model with state $x_k = [p_x, p_y, v_x, v_y]^T$ and apply the standard recursive estimator [27]:

$$\hat{x}_{k|k-1} = F\,\hat{x}_{k-1|k-1}\,, \quad P_{k|k-1} = F\,P_{k-1|k-1}\,F^T + Q \tag{1}$$

$$K_k = P_{k|k-1}\,H^T\,(H\,P_{k|k-1}\,H^T + R)^{-1} \tag{2}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\,(z_k - H\,\hat{x}_{k|k-1})\,, \quad P_{k|k} = (I - K_k\,H)\,P_{k|k-1} \tag{3}$$

where F encodes constant-velocity propagation over the sampling interval Δt, H selects the observed positional components, Q is the process-noise covariance reflecting the athlete's manoeuvring capability, and R is the measurement-noise covariance, which should be inflated when the reported dilution of precision degrades. Because the empirical corpus used in this study contains event annotations rather than raw positional telemetry, this component is specified and instantiated in the pipeline but is not empirically evaluated here; Section V-D states this limitation explicitly.

All coordinates are normalised to a common attacking-left-to-right frame of length L = 120 and width B = 80 units, so that for an event recorded by the team attacking in the negative direction:

$$(\tilde{x}, \tilde{y}) = (L - x, B - y) \tag{4}$$

3.4 Feature Construction

Layer L3 materialises four feature families from the conditioned stream. Geometric descriptors capture the relationship between the ball and the goal. With the goal centre at $g = (120, 40)$ and posts at $p_1 = (120, 36)$ and $p_2 = (120, 44)$, the shot distance and the angle subtended by the goal mouth at the shot location s are

$$d(s) = \|s - g\|_2\,, \quad \theta(s) = \arccos\left[\,(\|s-p_1\|^2 + \|s-p_2\|^2 - w^2) \,/\, (2\|s-p_1\| \cdot \|s-p_2\|)\,\right] \tag{5}$$

with w = 8 the goal width. Occupancy descriptors summarise the freeze frame. Let O be the set of opponent positions and T the triangle spanned by the shot location and the two posts. We compute the number of opponents obstructing the shooting cone, the number within radii of three and five units, the distance and lateral deviation of the goalkeeper, and an inverse-distance defensive density that acts as a discrete surrogate for continuous pitch-control formulations [12], [15]:

$$n_{\text{cone}} = |\{\, o \in O : o \in T \,\}|\,, \quad \rho(s) = \Sigma_{o \in O}\; 1 \,/\, (1 + \|s - o\|_2) \tag{6}$$

The goalkeeper deviation is the perpendicular distance of the keeper from the line joining the shot location to the goal centre, which distinguishes a well-positioned keeper from one caught out of line. Contextual descriptors record possession length, elapsed possession time, play pattern, body part, technique and pressure status. Sequence descriptors, used by the temporal module, encode each possession as an ordered tensor of action embeddings. Table II summarises the families.

TABLE II. FEATURE FAMILIES MATERIALISED IN THE SHARED FEATURE STORE.

Family	Representative members	Count	Consumers
Geometric	distance, subtended angle, log and inverse distance, longitudinal and lateral offset, distance × angle	8	xG module
Occupancy (freeze frame)	opponents in cone, opponents within 3 and 5 units, team-mates within 5 units, goalkeeper distance and lateral deviation, nearest opponent, inverse-distance density, cone density, goalkeeper gap	10	xG module, explanation layer
Contextual	possession length, possession passes, elapsed possession time, tempo, play pattern, body part, technique, shot type, under-pressure flag, assist geometry	18	xG module
Sequence	per-action type embedding, normalised coordinates, distance to goal, inter-event interval, team continuity flag, cumulative possession time	22 per step	temporal module

Family	Representative members	Count	Consumers
Workload	weighted action load, ball-carry distance, minutes, acute and chronic aggregates, coupled, uncoupled and exponentially weighted ratios	8	load monitor

3.5 Shot Valuation

The shot-valuation module estimates the conditional probability that a shot with feature vector x becomes a goal. Three estimator families are compared under an identical feature representation. The regularised logistic model is

$$P(y = 1 \mid x) = \sigma(\beta_0 + \beta^{T}x) , \quad \sigma(z) = 1 / (1 + e^{-z}) \tag{7}$$

fitted by penalised maximum likelihood. The random forest averages B unpruned trees grown on bootstrap resamples with random feature subsetting at each split [28]:

$$\hat{f}(x) = (1/B) \Sigma_{b=1}^{B} T_b(x ; \Theta_b) \tag{8}$$

The gradient-boosted ensemble minimises a regularised objective in which each additive tree is fitted to a second-order approximation of the loss [29]:

$$\mathcal{L} = \Sigma_i l(y_i, \hat{y}_i) + \Sigma_k \Omega(f_k) , \quad \Omega(f) = \gamma T + \tfrac{1}{2} \lambda \|w\|^2 \tag{9}$$

with l the logistic loss, T the number of leaves and w the leaf weights. Hyper-parameters were fixed a priori rather than tuned on the test partition: 600 trees of depth four with learning rate 0.05, subsampling of 0.8 for rows and columns, minimum child weight ten and L2 penalty two for the boosted model; 400 trees with a minimum leaf size of twenty and square-root feature subsetting for the forest.

3.6 Expected Threat Surface

The tactical valuation module estimates a zone-value surface over a 16 × 12 partition of the pitch, following the Markov formulation of Expected Threat [11]. For each zone z let s_z and m_z denote the empirical probabilities that an action originating in z is a shot or a move (pass, carry or dribble), g_z the conversion probability of shots taken from z, and T(z, z′) the probability that a move from z terminates successfully in z′, with the residual mass corresponding to turnover and absorbed at value zero. The zone value is the fixed point of

$$xT(z) \leftarrow s_z g_z + m_z \Sigma_{z'} T(z, z') xT(z') \tag{10}$$

iterated to convergence, and the value added by a successful move from z to z′ is

$$\Delta xT = xT(z') - xT(z) \tag{11}$$

Because the operator in (10) is a contraction whenever the expected turnover probability is strictly positive, the iteration converges geometrically from any non-negative initialisation. Only open-play actions are used, so that set-piece routines do not distort the transition structure.

3.7 Real-Time Threat Forecasting

The temporal module addresses the decision-support question that arises during live play: given the last few actions of the current possession, how likely is a shot in the immediate future? For each position i in a possession we construct a window of the L = 8 most recent actions, each represented by F = 22 features, and predict whether the possessing team registers a shot within H = 10 s. The temporal convolutional network applies dilated causal convolutions so that the receptive field grows exponentially with depth while no future information leaks into the prediction [30]. For layer ℓ with dilation d_ℓ and kernel width k:

$$h_\ell(t) = \mathrm{ReLU}(\Sigma_{j=0}^{k-1} W_\ell(j) h_{\ell-1}(t - d_\ell j) + b_\ell) + h_{\ell-1}(t) \tag{12}$$

with dilations d = (1, 2, 4), kernel width k = 3 and 32 channels, giving a receptive field of fifteen actions, which exceeds the window length so that the whole window is visible at the output step. The prediction is read from the terminal position of the final residual stack:

$$\hat{y} = \sigma(w^{T} h_{L}(t = L) + b) \tag{13}$$

and parameters are estimated by minimising the binary cross-entropy

$$\mathscr{L} = -(1/N) \Sigma_i [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \tag{14}$$

using Adam with a learning rate of 3×10^{-3}, batch size 1,024, for four epochs. Two reference models consume the identical window flattened to a 176-dimensional vector: a logistic regression and a gradient-boosted ensemble of 400 trees of depth six.

3.8 Workload Representation

The load module operates on a match-level external load index derived from event annotations. Let n_a be the count of actions of type a performed by a player in a fixture, w_a a type-specific intensity weight, and D_c the total ball-carrying distance. The event load is

$$EL = \Sigma_a w_a n_a + \alpha D_c , \quad \alpha = 0.05 \tag{15}$$

with weights of 1.5 for duels, dribbles and shots, 1.0 for pressures, interceptions, blocks, clearances and fouls, 0.8 for ball recoveries, 0.5 for miscontrols and dispossessions, and 0.3 for all remaining event types. Three ratio representations are computed over calendar windows. The conventional coupled ratio compares a seven-day acute aggregate with a twenty-eight-day chronic aggregate expressed on the same weekly scale:

$$ACWR_c = A_7 / (C_{28} / 4) \tag{16}$$

The uncoupled variant removes the acute window from the chronic term to eliminate the shared component identified as a source of spurious correlation [21]:

$$ACWR_u = A_7 / ((C_{28} - A_7) / 3) \tag{17}$$

and the exponentially weighted variant applies geometrically decaying weights with $\lambda_N = 2/(N+1)$ [18]:

$$EWMA_N(t) = \lambda_N EL(t) + (1 - \lambda_N) EWMA_N(t - 1) , \quad ACWR_e = EWMA_7 / EWMA_{28} \tag{18}$$

3.9 Evaluation Protocol

All partitions are formed at match level, never at event level, so that no possession contributes observations to both training and test sets. Twenty per cent of matches (303 of 1,517) were reserved as a held-out test set; the remaining 1,214 matches formed the development set, on which five-fold match-grouped cross-validation was performed. Discrimination is reported as the area under the receiver operating characteristic and the area under the precision-recall curve. Probabilistic quality is reported as logarithmic loss, the Brier score, the root-mean-square and mean absolute errors of the predicted probability against the binary outcome, and expected calibration error over ten equal-width bins, in which acc_b and conf_b denote the observed outcome frequency and the mean predicted probability within bin b:

$$BS = (1/N) \Sigma_i (\hat{y}_i - y_i)^2 , \quad ECE = \Sigma_{b=1}^{B} (N_b/N) \, | acc_b - conf_b | \tag{19}$$

Operational cost is reported as median and 95th-percentile single-event inference latency over 200 repetitions, batch throughput at a batch size of 1,024, serialised model size and training time. All measurements were taken on a single x86-64 core with 3 GB of RAM and no accelerator, which represents a conservative proxy for edge deployment on touchline hardware.

Attribution is computed with tree-based Shapley additive explanations, which provide exact per-observation feature contributions for tree ensembles and satisfy local accuracy and consistency [31]. Implementation used Python with scikit-learn 1.8.0, XGBoost 3.4.1, SHAP 0.52.0, JAX 0.11.1, NumPy 2.4.4 and pandas 3.0.2, with the global random seed fixed at 42.

4. EMPIRICAL RESULTS

4.1 Ingestion and Processing Throughput

The extraction stage retrieved and parsed all 1,517 match files in 87.8 s of wall-clock time using twelve concurrent I/O threads on a single core, corresponding to an end-to-end rate of 60,609 events per second including network transfer, deserialisation, feature construction and serialisation of the derived tables. Isolating parsing from network transfer on a locally cached subset of twenty matches gives 85,682 events per second on one core. Table III summarises the measured stage costs. For context, a live optical feed at 25 Hz for twenty-two players and the ball generates approximately 575 observations per second, roughly two orders of magnitude below the measured single-core conditioning capacity, so the ingestion path is not the binding constraint for real-time operation.

TABLE III. MEASURED PIPELINE COSTS (SINGLE CORE, 3 GB RAM, NO ACCELERATOR).

Stage	Scale	Wall-clock time	Throughput
Retrieval, parsing and feature extraction (12 I/O threads)	1,517 matches / 5,321,459 events	87.8 s	60,609 events s^{-1}
Parsing only, cached input, single thread	20 matches / 70,926 events	0.83 s	85,682 events s^{-1}
Expected Threat estimation, 68 iterations	2,360,390 open-play actions	10.8 s	—
Sequence tensor construction	2,238,130 windows	41.1 s	54,455 windows s^{-1}
xG model training (gradient boosting)	29,954 shots × 55 features	2.31 s	—
TCN training, 4 epochs	600,000 windows	33.7 s	71,200 windows s^{-1}

4.2 Shot Valuation: Comparative Evaluation

Table IV reports the comparative evaluation of analytic approaches on the shot-valuation task. Three findings stand out. First, the freeze-frame occupancy features are decisive: the purely geometric baseline using only distance and angle attains an AUC of 0.747, whereas every model with access to defensive configuration exceeds 0.815, an improvement of about seven points that dwarfs any difference between learning algorithms. Second, the choice of estimator family is close to immaterial for this task. Logistic regression, random forest and gradient boosting are separated by 0.002 AUC on the held-out matches, well within the cross-validation dispersion of ±0.007 to ±0.008. Third, and most important for deployment, the three models differ by orders of magnitude in operational cost. The random forest requires 31.9 MB of memory and 24.8 ms per single-event prediction, whereas the logistic model requires 5.8 kB and the boosted ensemble 0.88 MB, both at roughly 4 to 5 ms including feature-pipeline overhead. An architect optimising for the touchline would therefore select the simplest model, not the most flexible one.

The comparison against the commercial model distributed with the same corpus provides an external reference. The vendor model attains 0.819 AUC on the identical held-out matches, ahead of our best open model by 0.002 AUC and 0.0008 Brier score. We interpret this as parity rather than superiority in either direction: a fully reproducible open pipeline built from documented features matches a proprietary production model on this corpus.

TABLE IV. COMPREHENSIVE METHOD COMPARISON FOR SHOT VALUATION. TEST METRICS ON 303 HELD-OUT MATCHES (7,534 SHOTS, 772 GOALS). CV = FIVE-FOLD MATCH-GROUPED CROSS-VALIDATION ON 1,214 DEVELOPMENT MATCHES.

Analytic approach	Features	AUC (CV)	AUC (test)	PR-AUC	Latency p50 / p95 (ms)	Throughput (events s^{-1})	Size (MB)
Classical regression (distance, angle)	2 geometric	—	0.747	0.272	—	—	<0.01
Logistic regression	55 (all families)	0.811 ± 0.007	0.815	0.407	4.32 / 5.09	145,763	0.006
Random forest (400 trees)	55 (all families)	0.810 ± 0.007	0.816	0.399	24.78 / 29.93	12,136	31.94
XGBoost (600 trees, depth 4)	55 (all families)	0.806 ± 0.008	0.817	0.404	4.84 / 5.79	77,587	0.877
Vendor model (StatsBomb xG)	proprietary	—	0.819	0.418	n/a	n/a	n/a

Figure 3 shows the corresponding receiver operating characteristics and reliability diagrams. All models are well calibrated across the operating range, with mild over-confidence in the highest probability decile where sample sizes are smallest.

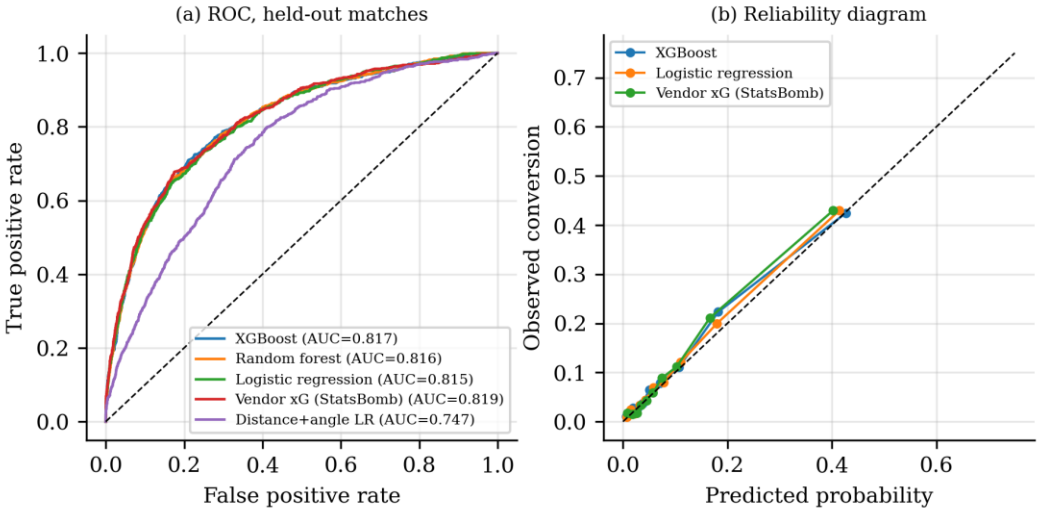

Fig. 3. (a) Receiver operating characteristics on 303 held-out matches. (b) Reliability diagrams with ten quantile bins; the diagonal denotes perfect calibration.

Table V reports the statistical validation metrics requested for operational monitoring. Aggregate accuracy is of particular practical interest because expected-goals totals are used to summarise performance over blocks of fixtures: all models under-predict the observed 772 goals in the test partition by between 6 and 8 %, reflecting the well-known excess of the observed conversion rate over the seasonal mean in any finite sample.

TABLE V. STATISTICAL VALIDATION OF SHOT-VALUATION MODELS ON THE HELD-OUT PARTITION. PRECISION, RECALL AND F1 ARE EVALUATED AT A DECISION THRESHOLD OF 0.30.

Model	RMSE	MAE	Log loss	Brier	ECE	Precision	Recall	Predicted goals
Distance + angle	0.2905	0.1617	0.2940	0.0844	0.0141	0.371	0.146	712.6
Logistic regression	0.2745	0.1458	0.2619	0.0754	0.0093	0.468	0.326	724.1
Random forest	0.2752	0.1480	0.2624	0.0758	0.0075	0.469	0.299	717.2
XGBoost	0.2748	0.1430	0.2623	0.0755	0.0153	0.458	0.328	709.3
Vendor model	0.2734	0.1442	0.2610	0.0747	0.0108	0.496	0.316	705.9
Observed	—	—	—	—	—	—	—	772

4.3 Feature Attribution

Figure 4 and Table VI report Shapley attributions for the boosted model over 4,000 held-out shots. The ranking is dominated by occupancy rather than geometry: the cone-density interaction, goalkeeper distance and nearest-opponent distance together account for a larger mean absolute contribution than the shot angle, and raw distance ranks only fifteenth. This is a substantively useful result for coaching staff, because it locates the controllable component of shot quality in the defensive configuration a team is able to create rather than in the location from which a shot is struck.

TABLE VI. TOP TEN FEATURES BY MEAN ABSOLUTE SHAP VALUE (LOG-ODDS UNITS), GRADIENT-BOOSTED XG MODEL.

#	Feature	Mean |SHAP|	#	Feature	Mean |SHAP|
1	Cone density (opponents in cone / angle)	0.513	6	Freeze-frame player count	0.097
2	Goalkeeper distance	0.402	7	Longitudinal offset to goal line	0.096

#	Feature	Mean |SHAP|	#	Feature	Mean |SHAP|
3	Nearest opponent distance	0.270	8	Goalkeeper gap (distance − keeper distance)	0.096
4	Subtended goal angle	0.244	9	Goalkeeper lateral deviation	0.094
5	Header indicator	0.123	10	Possession length	0.088

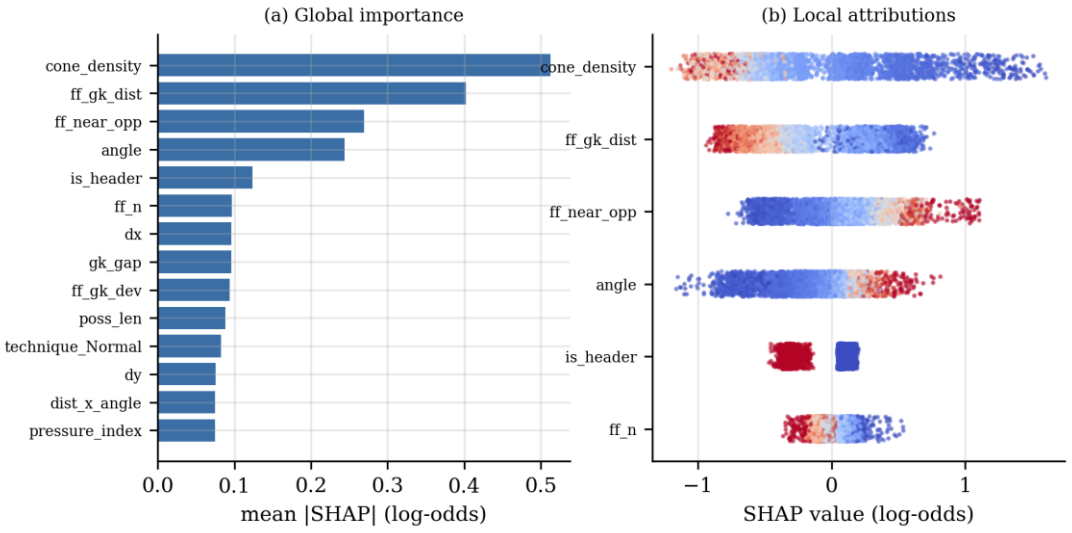

Fig. 4. (a) Global feature importance by mean absolute SHAP value. (b) Local attributions for the six highest-ranked features; colour encodes the normalised feature value from low (blue) to high (red).

4.4 Expected Threat Surface and Validation

The Markov recursion (10) was estimated on 2,360,390 open-play actions, of which 2,324,658 were moves and 35,732 were shots, with 2,057,966 moves terminating successfully. The iteration converged in 68 sweeps to a maximum absolute update of 9.1×10^{-8}, with the update norm falling from 0.265 after the first sweep to 5.8×10^{-4} after twelve. Table VII summarises the resulting surface, and Fig. 5 displays it.

TABLE VII. ESTIMATED EXPECTED THREAT SURFACE, 16 × 12 PARTITION.

Quantity	Value	Interpretation
Maximum zone value	0.2926	central six-yard zone
Mean over the pitch	0.0174	unconditional zone value
Mean, defensive third	0.0045	build-up origin
Mean, middle third	0.0095	progression corridor
Mean, final third	0.0398	8.9× the defensive third
Penalty-spot zones	0.2822	primary scoring region
Iterations to $\|\Delta\| < 10^{-7}$	68	geometric convergence

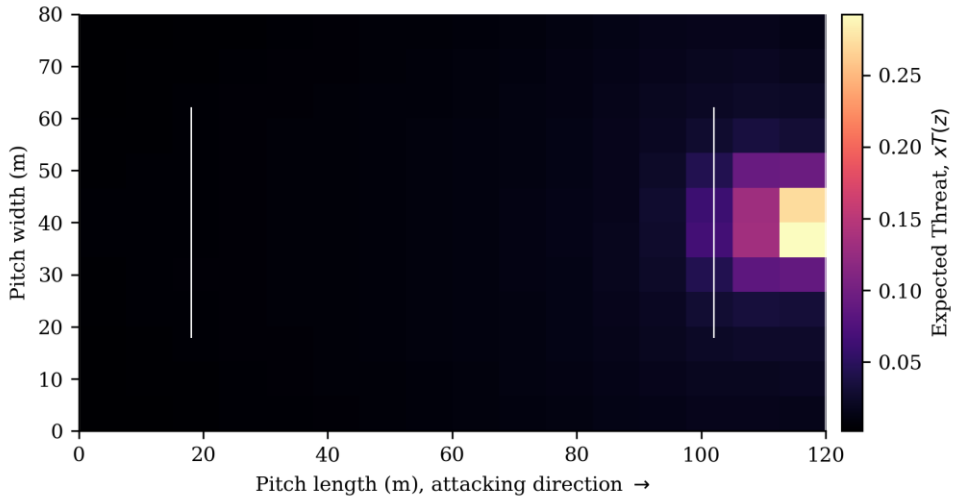

Fig. 5. Estimated Expected Threat surface over a 16 × 12 partition, 2015/16 four-league corpus. White lines mark the penalty areas; the attacking direction is left to right.

Because a value surface is only useful if it tracks outcomes it did not observe, we aggregated the per-action threat added by every successful open-play move to team-season totals and compared these with goals scored and league points, neither of which enters the estimation. Across 80 team-seasons the accumulated threat correlates with goals scored at $r = 0.859$ (Spearman $\rho = 0.789$, $p = 3.7 \times 10^{-18}$) and with league points at $r = 0.720$ ($\rho = 0.663$, $p = 2.0 \times 10^{-11}$). Table VIII lists the leading players by threat added per 90 minutes among those with at least 900 minutes; the ranking is dominated by wide creators and advanced playmakers, which is the expected signature of a metric that rewards progressive ball movement into high-value zones rather than finishing.

TABLE VIII. Leading players by Expected Threat added per 90 minutes (minimum 900 minutes played), 2015/16 four-league corpus.

#	Player	Team	Minutes	Actions	ΔxT per 90
1	Neymar da Silva Santos Junior	Barcelona	3,129	3,707	0.433
2	Gerard Deulofeu Lázaro	Everton	1,445	1,136	0.408
3	Lionel Andrés Messi Cuccittini	Barcelona	2,801	3,710	0.406
4	Ángel Fabián Di María Hernández	Paris Saint-Germain	2,068	2,711	0.403
5	Dries Mertens	Napoli	1,193	1,036	0.352
6	Dušan Tadić	Southampton	2,339	1,633	0.341
7	Gareth Frank Bale	Real Madrid	1,771	1,309	0.331
8	Manuel Agudo Durán	Celta Vigo	2,549	1,971	0.317
9	Cristian Tello Herrera	Fiorentina	1,092	913	0.315
10	Alexis Alejandro Sánchez Sánchez	Arsenal	2,510	2,828	0.315

4.5 Real-Time Threat Forecasting

The temporal module was evaluated on 2,238,130 sliding windows, of which 5.59 % are followed by a shot within ten seconds. Training used a random subsample of 600,000 windows from development matches and evaluation used 300,000 windows from held-out matches. Table IX reports the comparison. The temporal convolutional network attains an AUC of 0.802 against 0.805 for gradient boosting on the identical flattened windows and 0.755 for logistic regression, but it does so with 9,089 parameters occupying 36 kB, against 1.82 MB for the boosted ensemble. Its median single-window latency of 0.034 ms is 6.7 times lower than that of the boosted model, and its batch throughput is 2.5 times higher. For an embedded or touchline deployment in which the model must be co-resident with video processing on constrained hardware, this trade-off favours the convolutional architecture decisively; for a server-side deployment with no memory pressure, the boosted model remains a reasonable default.

TABLE IX. REAL-TIME THREAT FORECASTING: SHOT WITHIN A TEN-SECOND HORIZON. 300,000 HELD-OUT WINDOWS, POSITIVE RATE 5.59 %. PRECISION, RECALL AND F1 AT A THRESHOLD OF 0.20.

Model	AUC	PR-AUC	Brier	ECE	Latency p50 (ms)	Throughput (s^{-1})	Parameters	Size (MB)
Logistic regression (flattened window)	0.755	0.154	0.0500	0.0030	0.055	3,676,276	177	<0.01
XGBoost (flattened window)	0.805	0.213	0.0480	0.0012	0.226	157,977	400 trees	1.82
TCN (3 dilated blocks, 32 channels)	0.802	0.207	0.0485	0.0121	0.034	389,083	9,089	0.036

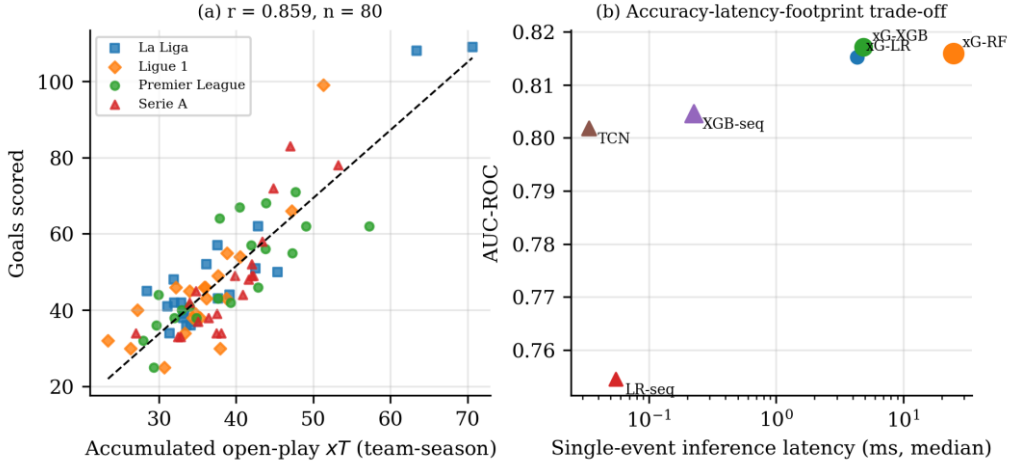

Fig. 6. (a) Team-season Expected Threat against goals scored, 80 team-seasons. (b) Accuracy against single-event latency for all six models; marker area scales with serialised model size.

4.6 Workload Monitoring

The load module was computed over 42,093 player-fixture records covering 2,176 players. The event-derived load index correlates with minutes played at r = 0.669 (n = 42,093), confirming that it captures exposure volume while retaining independent variation attributable to intensity: mean load per 90 minutes is 82.5 ± 30.9 units. Ratio representations were computed for the 34,003 player-fixtures preceded by at least four appearances.

Table X exposes a structural property of match-only load data that is important for practitioners. The conventional coupled ratio is severely inflated in this setting: its mean is 1.68 and 48.2 % of player-fixtures exceed the conventional 1.5 threshold, because a footballer who plays one fixture in a twenty-eight-day window has an acute aggregate equal to the entire chronic aggregate, driving the ratio to its structural maximum of four. The exponentially weighted representation is far better behaved, with a mean of 1.07, a standard deviation of 0.26, and only 6.7 % of observations above 1.5. The coupled and uncoupled ratios correlate at r = 0.935 with each other but at only r = 0.031 with the exponentially weighted variant, so the two families are not interchangeable and should not be compared across studies. These observations are consistent with the methodological critique of ratio-based load indices [19]–[21] and support the recommendation to prefer decaying weights [18].

TABLE X. DISTRIBUTION OF WORKLOAD RATIO REPRESENTATIONS OVER 34,003 PLAYER-FIXTURES.

Representation	Mean	SD	Median	95th pct.	% above 1.5	% below 0.8
Coupled 7:28 rolling	1.676	0.922	1.465	4.000	48.2 %	12.5 %
Uncoupled 7:21 rolling	2.335	2.136	1.612	7.418	53.9 %	15.8 %
EWMA (λ_7, λ_{28})	1.069	0.259	1.024	1.578	6.7 %	8.3 %

No open dataset pairs these load representations with injury labels, so we do not claim injury prediction. We instead evaluate an availability proxy: whether a player who featured in a fixture also features in the club's next fixture. Across 32,932 observations the base non-participation rate is 19.3 %. Table XI shows a clear U-shaped relationship. Players in the

conventional sweet spot are least likely to be absent, players above 1.5 are absent at 20.6 % (odds ratio 1.42 against the 0.8–1.3 reference band, $\chi^2 = 103.2$, $p = 3.0 \times 10^{-24}$), and players below 0.8 are absent at 26.2 % — the highest rate in the table. The low-ratio band is dominated by players returning from limited exposure, so the association there is at least partly reverse causation. A temporally split logistic model using workload features reaches an AUC of 0.668 on later fixtures against 0.619 for a model using only minutes played and rest days, an improvement of 0.049 attributable to the load representations.

TABLE XI. NON-PARTICIPATION IN THE NEXT FIXTURE BY COUPLED WORKLOAD RATIO BAND (32,932 OBSERVATIONS).

ACWR band	Observations	Non-participation in next fixture	Mean minutes played	Mean event load
< 0.8	4,032	26.2 %	50.3	37.6
0.8 – 1.0	3,622	14.8 %	79.5	65.8
1.0 – 1.3	5,733	15.8 %	81.5	72.4
1.3 – 1.5	3,488	15.0 %	82.4	75.5
> 1.5	16,030	20.6 %	77.8	72.6

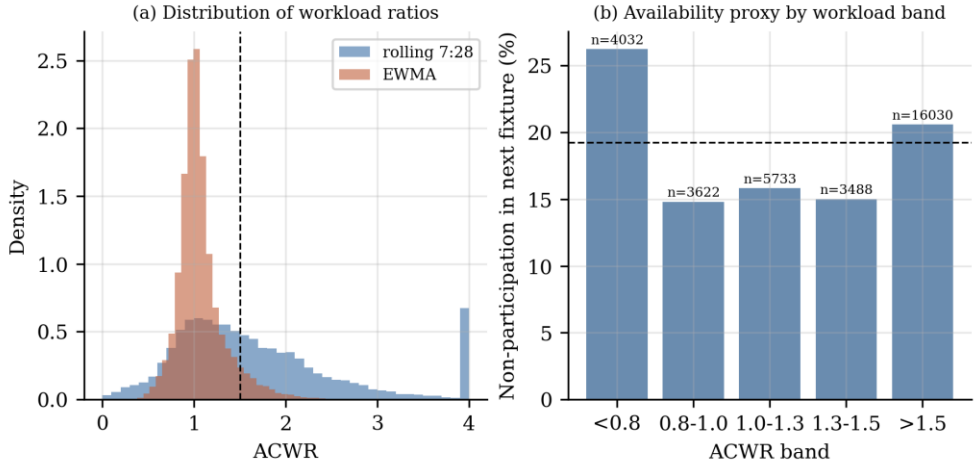

Fig. 7. (a) Distribution of coupled rolling and exponentially weighted workload ratios; the dashed line marks the conventional 1.5 threshold. (b) Non-participation rate by workload band; the dashed line marks the base rate.

4.7 Cross-Corpus Generalisation

A model validated only on the season it was fitted to tells a coaching staff nothing about whether it will still work next year, or whether it can be deployed in another competition. We therefore assembled two further corpora from the same open collection and scored them with the frozen 2015/16 models, without refitting, feature reselection or hyper-parameter adjustment. Corpus B covers modern men's football: seven competitions played between 2020 and 2024, including three continental championships and a World Cup. Corpus C covers women's football: six competitions played in 2023 and 2023/24 across five countries. Together they add 1,135 matches and 28,883 shots, none of which contributed any observation to model fitting. Shot freeze frames are present for 100 % of shots in both corpora, so the occupancy features are fully available. Table XII describes the composition.

TABLE XII. COMPOSITION OF THE EXTERNAL VALIDATION CORPORA. NO OBSERVATION FROM EITHER CORPUS WAS USED IN FITTING, FEATURE SELECTION OR HYPER-PARAMETER CHOICE.

Corpus	Competition	Matches	Shots*	Goals*	Conversion
B	FIFA World Cup 2022	64	1,430	152	10.6 %
B	UEFA Euro 2024	51	1,304	98	7.5 %
B	Africa Cup of Nations 2023	52	1,162	98	8.4 %
B	Copa América 2024	32	741	63	8.5 %

Corpus	Competition	Matches	Shots*	Goals*	Conversion
B	Bundesliga 2023/24	34	908	102	11.2 %
B	Ligue 1 2022/23	32	840	102	12.1 %
B	La Liga 2020/21	35	827	102	12.3 %
B	Subtotal (modern men's football)	300	7,212	717	9.9 %
C	Liga F 2023/24	240	6,114	701	11.5 %
C	Serie A Women 2023/24	130	3,687	355	9.6 %
C	FA Women's Super League 2023/24	132	3,506	402	11.5 %
C	NWSL 2023	137	3,490	299	8.6 %
C	Frauen-Bundesliga 2023/24	132	3,266	380	11.6 %
C	FIFA Women's World Cup 2023	64	1,608	136	8.5 %
C	Subtotal (women's football)	835	21,671	2,273	10.5 %
B + C	Total	1,135	28,883	2,990	10.4 %

*Excluding penalty kicks.

Table XIII reports the outcome and Fig. 8 displays it. Four findings emerge. First, discrimination degrades under transfer, but modestly and symmetrically: the boosted model falls from 0.817 in domain to 0.799 on modern men's football and 0.796 on women's football. Critically, the commercial reference model degrades by almost exactly the same amount over the same shots, from 0.819 to 0.805 and 0.800. The gap between our open model and the vendor model therefore remains between 0.004 and 0.007 in all three corpora. The loss is a property of the target corpora — tournament football has a different chance profile from league football, and annotation batches differ across releases — rather than a failure of the transferred model.

Second, calibration degrades more than discrimination, and unevenly across estimator families. The boosted model transfers with a calibration slope of 0.88 to 0.89, indicating predictions that are slightly too extreme off-domain, whereas the random forest transfers at 0.99 to 1.01 and the logistic model at 0.93 to 0.94. The flexible model has fitted more of the source distribution and pays for it when that distribution shifts. Because the deficiency is a single-parameter one, it is also cheap to repair: a one-parameter logistic recalibration on a few hundred target-corpus shots restores the slope without retraining.

Third, aggregate bias is the practically consequential failure mode. The frozen boosted model under-predicts goals by 3.3 % on modern men's football but by 10.1 % on women's football. Any club or federation using a men's-trained expected-goals model to produce team reports in the women's game will therefore systematically undervalue the chances created, by roughly a tenth, unless the model is recalibrated. We regard this as the single most actionable result of the transfer study.

Fourth, and least expected, refitting on the target corpus is not automatically better than transferring. Refitting the same architecture on corpus B under five-fold match-grouped cross-validation yields an AUC of 0.754, substantially below the 0.799 achieved by the model trained on the larger out-of-domain corpus, because 7,212 shots are not enough to estimate a 55-feature ensemble. On corpus C, where 21,671 shots approach the size of the source training set, refitting reaches 0.796 — statistically indistinguishable from transfer — but improves aggregate bias markedly, from −10.1 % to −1.5 %. The operational reading is that training-set volume dominates domain match for discrimination, while domain match dominates for calibration.

TABLE XIII. CROSS-CORPUS GENERALISATION OF SHOT VALUATION. FROZEN MODELS WERE FITTED ON THE 2015/16 DEVELOPMENT MATCHES AND APPLIED WITHOUT ADJUSTMENT. THE REFITTED ROW IS FIVE-FOLD MATCH-GROUPED CROSS-VALIDATION WITHIN THE TARGET CORPUS.

Corpus	Model	AUC	PR-AUC	Brier	ECE	Calibration slope	Goal bias (%)
A: 2015/16 held-out (in domain)	Logistic regression	0.815	0.407	0.0754	0.0093	0.978	−6.2

Corpus	Model	AUC	PR-AUC	Brier	ECE	Calibration slope	Goal bias (%)
	Random forest	0.816	0.399	0.0758	0.0075	1.031	−7.1
	XGBoost	0.817	0.404	0.0755	0.0153	0.914	−8.1
	Vendor model	0.819	0.418	0.0747	0.0108	1.025	−8.6
B: modern men's, 2020–2024 (7,212 shots)	Logistic regression	0.798	0.376	0.0754	0.0092	0.938	−3.5
	Random forest	0.798	0.370	0.0756	0.0032	1.014	−1.9
	XGBoost (frozen)	0.799	0.365	0.0760	0.0125	0.881	−3.3
	Vendor model	0.805	0.395	0.0742	0.0066	1.012	−2.1
	XGBoost refitted on corpus B	0.754	0.312	0.0814	0.0288	0.653	−4.0
C: women's, 2023/24 (21,671 shots)	Logistic regression	0.796	0.397	0.0780	0.0097	0.933	−8.6
	Random forest	0.793	0.390	0.0785	0.0079	0.996	−6.2
	XGBoost (frozen)	0.796	0.398	0.0780	0.0115	0.890	−10.1
	Vendor model	0.800	0.416	0.0769	0.0068	1.005	−5.9
	XGBoost refitted on corpus C	0.796	0.397	0.0780	0.0091	0.864	−1.5

The Expected Threat surface behaves altogether differently. Re-estimating (10) independently on each validation corpus — 520,033 open-play moves for corpus B and 1,252,688 for corpus C — produces surfaces that are nearly identical to the 2015/16 surface. The zone-value correlation with the source surface is r = 0.992 for modern men's football and r = 0.997 for women's football, with rank correlations of 0.999 in both cases and mean absolute zone differences of 0.0032 and 0.0014. Maximum zone values are 0.287 and 0.278 against 0.293 in the source corpus. Table XIV summarises the comparison and Fig. 8(b) shows the agreement zone by zone.

TABLE XIV. EXPECTED THREAT SURFACES RE-ESTIMATED INDEPENDENTLY ON EACH CORPUS, COMPARED WITH THE 2015/16 SURFACE.

Corpus	Open-play moves	Iterations	Maximum xT	Pearson / Spearman r vs. corpus A	Mean |difference|
A: 2015/16 four leagues	2,324,658	68	0.2926	—	—
B: modern men's	520,033	93	0.2867	0.992 / 0.999	0.0032
C: women's	1,252,688	69	0.2779	0.997 / 0.999	0.0014

This dissociation is the central methodological result of the transfer study. The spatial value geometry of the pitch — how much more threatening one zone is than another — is close to invariant across an eight-year interval and across the men's and women's games, so a threat surface estimated once can be reused across competitions with little loss. The conditional finishing model is not invariant: it transfers with acceptable discrimination but biased magnitude, and requires recalibration before its outputs are aggregated. A deployment that reuses both artefacts uncritically would inherit a systematic error in exactly the quantity practitioners report most often.

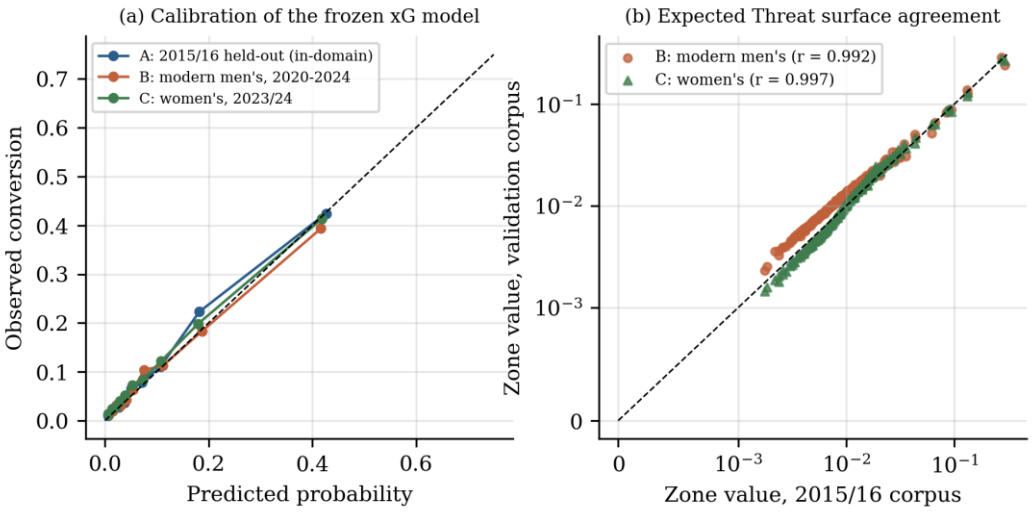

Fig. 8. (a) Reliability of the frozen gradient-boosted expected-goals model on the in-domain held-out partition and on the two external corpora. (b) Zone-by-zone agreement between the Expected Threat surface estimated on the 2015/16 corpus and surfaces estimated independently on each validation corpus; axes are symmetric-log and the dashed line is the identity.

5. DISCUSSION

5.1 Translating Model Output into Interventions

The results support three concrete classes of intervention. In tactical preparation, the dominance of occupancy features over geometry in Table VI implies that chance-quality improvement is better pursued through actions that alter defensive configuration — attacking the space behind a compressed line, or engineering cut-backs that leave no defender inside the shooting cone — than through shot selection rules based on distance alone. The Expected Threat surface makes this quantitative: the mean value of the final third exceeds that of the defensive third by a factor of 8.9, so possession sequences can be ranked by the threat they generate per turnover risk incurred, and specific carrying or passing lanes can be prioritised in training.

In live match support, the temporal module produces a continuously updated probability that the current possession will produce a shot within ten seconds. At a latency of 0.034 ms per window, the module can be evaluated for both teams on every incoming action with a budget three orders of magnitude below the inter-event interval, leaving ample headroom for the video and tracking stack that would co-reside on the same hardware. Practically, the useful output is not the raw probability but its aggregation: sustained divergence between the threat a team generates and the threat it concedes during a match phase is an actionable signal for a pressing or block adjustment.

In physical preparation, the module supports rotation decisions rather than injury prediction. The appropriate use is to flag players whose exponentially weighted ratio has moved outside its individual historical range, to present that flag alongside minutes and rest days, and to leave the decision to clinical staff. The instability of the coupled ratio documented in Table X is a direct argument against automated threshold alerts built on that representation.

5.2 Interpretability for Practitioner Adoption

Shapley attributions serve two distinct purposes in this framework. Globally, they support model validation by domain experts: an analyst can confirm that the model has learned that headers convert at lower rates, that goalkeeper positioning matters, and that possession length has only a marginal contribution. Locally, they convert an individual prediction into a sentence a coach can act on — for example, that a particular chance was valued at 0.31 principally because two defenders were removed from the shooting cone and the goalkeeper was 4 m off his line. Attribution computation for 4,000 shots required 3.1 s, so per-chance explanations are available at post-match reporting timescales without special infrastructure.

5.3 Edge Cases and Data Pathologies

Three failure modes dominate operational deployments. Missing tracking frames occur when players are occluded or leave the camera envelope; the state-space formulation in (1)–(3) handles this natively by propagating the prediction step without a measurement update and allowing the covariance to grow, which both interpolates the gap and marks the interpolated interval as low-confidence for downstream consumers. Features that depend on complete configuration, such as the cone count in (6), should be suppressed rather than imputed when the covariance exceeds a threshold, because a silently wrong occupancy count is more damaging than a missing one. In our corpus the analogous pathology is partial freeze frames: annotators record only players visible in frame, so the occupancy features are conditioned on visibility. We retain the freeze-frame player count as an explicit feature, and it ranks sixth in Table VI, indicating that the model uses annotation completeness as context rather than being misled by it.

Sensor drift in satellite systems manifests as slowly varying bias rather than white noise, and is not removed by the filter in (1)–(3) under a constant-velocity model. The practical mitigation is periodic re-anchoring against fixed pitch landmarks and monitoring of the innovation sequence: a persistently non-zero mean innovation is a drift signature and should trigger recalibration rather than silent correction. Noise during high-velocity phases is the third pathology, and is precisely the regime in which validity studies report the largest errors [3], [4]; inflating the measurement covariance R as a function of estimated speed is the standard remedy, at the cost of increased smoothing lag during exactly the transitions that tactical analysis cares most about. This tension is intrinsic and should be resolved explicitly per use case rather than by a single global filter setting.

5.4 Scope of Empirical Validation and Limitations

We state precisely which claims this study supports. Table XII maps each framework component to its evidential status. Three limitations deserve emphasis. First, the corpus is event data, not tracking data: the conditioning layer, including the Kalman formulation, is specified and implemented but is not empirically evaluated here, and the occupancy features derive from discrete freeze frames rather than continuous trajectories. Second, the availability analysis in Section IV-F is a proxy analysis. Non-participation in the next fixture conflates injury with rotation, tactical selection and suspension; the reported odds ratio and AUC are therefore evidence that the workload representations carry squad-availability signal, and are not evidence of injury prediction. We make no injury-prediction claim, and we regard any such claim built on open data alone as unwarranted. Third, the transfer study in Section IV-G covers shot valuation and threat surfaces only. The temporal forecasting module and the workload module were evaluated in domain alone, and their behaviour under the era and domain shifts documented for the xG model is unknown; we would expect the sequence model, which encodes tempo and possession structure, to be more sensitive to tactical drift than the threat surface proved to be.

The distributed processing layer is likewise specified rather than benchmarked at scale. The throughput figures in Table III are single-node measurements; they establish that the per-event cost is compatible with real-time operation, but they do not establish the scaling behaviour of the streaming deployment across a cluster, which would require a multi-node experiment we did not perform.

TABLE XV. EVIDENTIAL STATUS OF EACH FRAMEWORK COMPONENT IN THE PRESENT STUDY.

Component	Status	Basis
Ingestion and parsing (L1)	Evaluated	1,517 matches, 5,321,459 events, measured throughput
Kalman conditioning (L2)	Specified only	Corpus contains event annotations, not positional telemetry
Distributed streaming (L2)	Specified only	Single-node throughput measured; cluster scaling not tested
Feature store (L3)	Evaluated	55 shot features, 22-dimensional sequence steps, workload indices
Shot valuation (L4)	Evaluated	Match-level hold-out and 5-fold grouped CV, vendor benchmark
Expected Threat (L4)	Evaluated	Convergence analysis and team-season external validation
Threat forecasting (L4)	Evaluated	300,000 held-out windows, three architectures
Workload monitoring (L4)	Partially evaluated	Distributions computed; availability proxy only, no injury labels
Injury risk prediction	Not evaluated	No open corpus pairs workload with injury outcomes

Component	Status	Basis
Explanation layer (L5)	Evaluated	Exact tree SHAP over 4,000 held-out shots
External validation, shot valuation and xT	Evaluated	1,135 further matches and 28,883 shots across 13 competitions, frozen models
External validation, temporal and load modules	Not evaluated	Transfer study covers shot valuation and threat surfaces only

6. Conclusion and Future Work

This paper specified a five-layer big data analytics framework for decision support in football and evaluated it end to end on a fully public corpus of 1,517 matches and 5,321,459 events. The evaluation produced four findings with practical consequences. Spatial occupancy features derived from shot freeze frames are worth roughly seven points of AUC over geometric baselines, while the choice among logistic regression, random forests and gradient boosting is worth about two thousandths; an open pipeline built on documented features matches the proprietary model distributed with the same data. A compact temporal convolutional network matches gradient boosting on live threat forecasting with a fiftyfold smaller footprint and sub-0.04 ms latency, which makes edge deployment feasible. A zone-based threat surface estimated by fixed-point iteration converges in 68 sweeps and explains 74 % of the variance in team-season goals scored despite never observing a match result. And conventional coupled workload ratios are structurally unstable on match-only exposure data, placing 48 % of player-fixtures above the conventional risk threshold, whereas exponentially weighted ratios place 6.7 % there.

A fifth finding came from the external validation on 1,135 further matches. Value surfaces and finishing models generalise differently: threat surfaces re-estimated on modern men's and on women's football agree with the 2015/16 surface at $r \geq 0.992$, whereas the frozen finishing model, although it loses only 0.02 AUC — the same loss incurred by the commercial reference over the same shots — under-predicts goals in the women's game by 10.1 %. Practitioners can therefore port a threat surface between competitions but must recalibrate an expected-goals model before aggregating its output. A further consequence is that a large out-of-domain training set beat a small in-domain one outright: refitting on 7,212 target-corpus shots produced an AUC of 0.754 against 0.799 for the transferred model.

The broader methodological claim is that reporting accuracy without computational cost is insufficient for a decision-support literature. Two of the models evaluated here differ by a factor of 5,500 in memory and a factor of five in latency while differing by 0.001 in AUC; only one of them can be deployed on constrained hardware, and no accuracy-only comparison would reveal that.

Four directions follow. Integration with continuous tracking data would allow the conditioning layer to be validated empirically and would replace discrete occupancy counts with continuous pitch-control surfaces. Quantised deployment of the temporal module on wearable-class hardware is a natural extension given its 36 kB footprint, and would enable on-body inference without a telemetry round trip. Pose estimation from broadcast video would supply body orientation and shooting posture, which are absent from the current feature store and are plausibly a substantial part of the residual variance in shot outcomes. Finally, validating the workload module against clinically adjudicated injury outcomes within a club environment, under a pre-registered protocol with external validation, remains the necessary step before any load-based component of this framework should influence medical decisions.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Acknowledgment

None.

References

[1] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science," SpringerPlus, vol. 5, no. 1, Art. no. 1410, 2016, doi: 10.1186/s40064-016-3108-2.

[2] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," ACM Comput. Surv., vol. 50, no. 2, Art. no. 22, pp. 1–34, 2017, doi: 10.1145/3054132.

[3] R. J. Aughey, "Applications of GPS technologies to field sports," Int. J. Sports Physiol. Perform., vol. 6, no. 3, pp. 295–310, 2011, doi: 10.1123/ijspp.6.3.295.

[4] C. Cummins, R. Orr, H. O'Connor, and C. West, "Global positioning systems (GPS) and microtechnology sensors in team sports: A systematic review," Sports Med., vol. 43, no. 10, pp. 1025–1042, 2013, doi: 10.1007/s40279-013-0069-2.

[5] F. R. Goes et al., "Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review," Eur. J. Sport Sci., vol. 21, no. 4, pp. 481–496, 2021, doi: 10.1080/17461391.2020.1747552.

[6] M. Herold, F. Goes, S. Nopp, P. Bauer, C. Thompson, and T. Meyer, "Machine learning in men's professional football: Current applications and future directions for improving attacking play," Int. J. Sports Sci. Coach., vol. 14, no. 6, pp. 798–817, 2019, doi: 10.1177/1747954119879350.

[7] M. Rico-González, J. Pino-Ortega, A. Méndez, F. M. Clemente, and A. Baca, "Machine learning application in soccer: A systematic review," Biol. Sport, vol. 40, no. 1, pp. 249–263, 2023, doi: 10.5114/biolsport.2023.112970.

[8] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," Sci. Data, vol. 6, Art. no. 236, 2019, doi: 10.1038/s41597-019-0247-7.

[9] StatsBomb, "StatsBomb open data." [Online]. Available: https://github.com/statsbomb/open-data. Accessed: Sep. 4, 2026.

[10] T. Decroos, L. Bransen, J. Van Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD '19), Anchorage, AK, USA, 2019, pp. 1851–1861, doi: 10.1145/3292500.3330758.

[11] K. Singh, "Introducing Expected Threat (xT): Modelling team behaviour in possession to gain a deeper understanding of buildup play," 2019. [Online]. Available: https://karun.in/blog/expected-threat.html. Accessed: Sep. 4, 2026.

[12] J. Fernández, L. Bornn, and D. Cervone, "A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions," Mach. Learn., vol. 110, no. 6, pp. 1389–1427, 2021, doi: 10.1007/s10994-021-05989-6.

[13] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry, "A multiresolution stochastic process model for predicting basketball possession outcomes," J. Amer. Stat. Assoc., vol. 111, no. 514, pp. 585–599, 2016, doi: 10.1080/01621459.2016.1141685.

[14] G. Anzer and P. Bauer, "A goal scoring probability model for shots based on synchronized positional and event data in football (soccer)," Front. Sports Active Living, vol. 3, Art. no. 624475, 2021, doi: 10.3389/fspor.2021.624475.

[15] D. Link, S. Lang, and P. Seidenschwarz, "Real time quantification of dangerousity in football using spatiotemporal tracking data," PLoS ONE, vol. 11, no. 12, Art. no. e0168768, 2016, doi: 10.1371/journal.pone.0168768.

[16] T. J. Gabbett, "The training–injury prevention paradox: Should athletes be training smarter and harder?," Br. J. Sports Med., vol. 50, no. 5, pp. 273–280, 2016, doi: 10.1136/bjsports-2015-095788.

[17] S. Malone, A. Owen, M. Newton, B. Mendes, K. D. Collins, and T. J. Gabbett, "The acute workload ratio in relation to injury risk in professional soccer," J. Sci. Med. Sport, vol. 20, no. 6, pp. 561–565, 2017, doi: 10.1016/j.jsams.2016.10.014.

[18] S. Williams, S. West, M. J. Cross, and K. A. Stokes, "Better way to determine the acute workload ratio?," Br. J. Sports Med., vol. 51, no. 3, pp. 209–210, 2017, doi: 10.1136/bjsports-2016-096589.

[19] F. M. Impellizzeri, M. S. Tenan, T. Kempton, A. Novak, and A. J. Coutts, "Acute workload ratio: Conceptual issues and fundamental pitfalls," Int. J. Sports Physiol. Perform., vol. 15, no. 6, pp. 907–913, 2020, doi: 10.1123/ijspp.2019-0864.

[20] F. M. Impellizzeri, S. Woodcock, A. J. Coutts, M. Fanchini, A. McCall, and A. D. Vigotsky, "What role do chronic workloads play in the acute to chronic workload ratio? Time to dismiss ACWR and its underlying theory," Sports Med., vol. 51, no. 3, pp. 581–592, 2021, doi: 10.1007/s40279-020-01378-6.

[21] L. Lolli, A. M. Batterham, R. Hawkins, D. M. Kelly, A. J. Strudwick, R. T. Thorpe, W. Gregson, and G. Atkinson, "Mathematical coupling causes spurious correlation within the conventional acute-to-chronic workload ratio calculations," Br. J. Sports Med., vol. 53, p. 921, 2019, doi: 10.1136/bjsports-2017-098110.

[22] P. C. Bourdon et al., "Monitoring athlete training loads: Consensus statement," Int. J. Sports Physiol. Perform., vol. 12, no. s2, pp. S2-161–S2-170, 2017, doi: 10.1123/IJSPP.2017-0208.

[23] A. Rossi, L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernández, and D. Medina, "Effective injury forecasting in soccer with GPS training data and machine learning," PLoS ONE, vol. 13, no. 7, Art. no. e0201264, 2018, doi: 10.1371/journal.pone.0201264.

[24] H. Van Eetvelde, L. D. Mendonça, C. Ley, R. Seil, and T. Tischer, "Machine learning methods in sport injury prediction and prevention: A systematic review," J. Exp. Orthop., vol. 8, no. 1, Art. no. 27, 2021, doi: 10.1186/s40634-021-00346-x.

[25] M. Zaharia et al., "Apache Spark: A unified engine for big data processing," Commun. ACM, vol. 59, no. 11, pp. 56–65, 2016, doi: 10.1145/2934664.

[26] Apache Software Foundation, "Apache Kafka documentation." [Online]. Available: https://kafka.apache.org/documentation/. Accessed: Sep. 4, 2026.

[27] R. E. Kalman, "A new approach to linear filtering and prediction problems," J. Basic Eng., vol. 82, no. 1, pp. 35–45, 1960, doi: 10.1115/1.3662552.

[28] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[30] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271, 2018, doi: 10.48550/arXiv.1803.01271.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.