Defending the Long Throw: A Dual-Model Framework for Minimising Defensive Risk

Sebastian Greenhalgh, Dan Anghel, Saba Kutibashvili, Hugo Arsénio, Ramiro Pascual

1. Introduction

Long throw-ins have re-emerged as a deliberate attacking weapon in elite football. On the opening weekend of the 2025-26 Premier League season, 11 of 20 teams attempted at least one long throw, no doubt as a result of their success in the previous campaign. In fact, the previous season saw 19 goals being scored directly from throw-in situations thus representing the highest total of the last decade. Moreover, they are no longer considered as a tool for teams struggling to generate chances from open play, possession-dominant teams now exploit them as structured set pieces: loading the box in ways that replicate many of the aerial and second-ball dynamics of corners. Thus teams have realised long throw-ins represent an under-utilised opportunity for marginal gains.

Even though throw-ins remain the most frequent restart in football (approx. 40 per match), they remain under-analysed in existing literature. The research that does exist focuses primarily on attacking mechanics:

- Stone, Smith & Barry (2021) analysed all 16,154 Premier League throw-ins in 2018/19 and found that backward or lateral throws were associated with higher first-contact success and possession retention.[1]

- Casal et al. (2023) examined 2,658 LaLiga throw-ins. They identified that duration, pressure, distance, and zone were the key indicators of attacking success.[2]

Nevertheless, throw-in defensive behaviours remain largely unexplored and long throw-in defences even less so. The few studies that have attempted to do so tend to have inferred defensive recommendations indirectly, rather than testing specific structures, player profiles, or goalkeeper starting depths. Even Casal et al. explicitly noted the scarcity of multivariate, context-aware analyses of defensive setups.

This paper aims to address that gap directly. Using multi-season Hudl StatsBomb event and 360 data from the Premier League, Championship, and Bundesliga, we quantify how throw-in origin,

[1] Stone, J. A, Smith, A., & Barry, A. (2021) 'The undervalued set piece: Analysis of soccer throw-ins during the English Premier League 2018-2019 season.' International Journal of Sports Science & Coaching

[2] Casal et al. (2023) 'Effects of tactical dimension and situational variables in throw-ins on the offensive performance in football'

landing zone, crowding, and physical match-ups shape the likelihood of a shot or goal within a short timeframe after the throw and then use these metrics to link event outcomes to measurable tactical choices such as defensive line depth, marking density, and more.

Methodologically, we adopt two complementary approaches:

1. A feature-driven interpretable model (XGBoost). This isolates the geometric and physical factors most associated with shot creation and provides actionable insight for coaches on positioning and match-ups.
2. A temporal GRU sequence model. This captures the dynamic evolution of post-throw events, predicting shot and goal probabilities over time from sequential event data. The model then quantifies how defensive interactions evolve across passes, duels, and recoveries after the throw-in using probabilistic outputs to characterise the temporal risk profile.

The combination of these two models bridge interpretability and temporal prediction. This facilitates both a strategic diagnosis (why danger arises) and risk forecasting (when danger arises) and represents a contribution to a defensively focused framework for evaluating long throw-ins.

2. Statistical Analysis

2.1 Operational definitions and baseline outcomes

We define long throw-ins as a throw-in satisfying the following criteria:

1. Has a recorded pass length exceeding 20m.
2. Whose end location lies inside the opposition penalty area.

For the following section, we measure outcomes in a 15s window following the throw. However, this timeframe extends to 45s for corners (since the time associated with a corner is when it is taken, not when it is won). Labels are then mutually exclusive and prioritised as follows:

1. Goal (overrides shot)
2. Shot
3. Corner won
4. No outcome

Across five competition seasons - Premier League (2024-25; 2023-24), Championship (2024-25; 2023-24), Bundesliga (2024-25) - we identify 3,959 qualifying long throws.

- No outcome: 3,061 (77.51%)
- Shot: 581 (14.68%)
- Corner won: 245 (6.19%)
- Goal: 72 (1.82%)

This means that around three in four long throws yield no tangible outcome within 15s, while around one in six generate a shot or goal.

2.2 Spatial distribution and symmetry

Long-throw entry points cluster heavily within the half of the box where the throw originates from, particularly the zones between the six-yard line and penalty spot - indicating the physical limitations of throwers. Figures 2.1 and 2.2 show throw locations from the top and bottom touchlines respectively, revealing almost perfect bilateral symmetry. This confirms that outcomes are largely independent of the throw-in side, thus henceforth, results combine both touchlines.

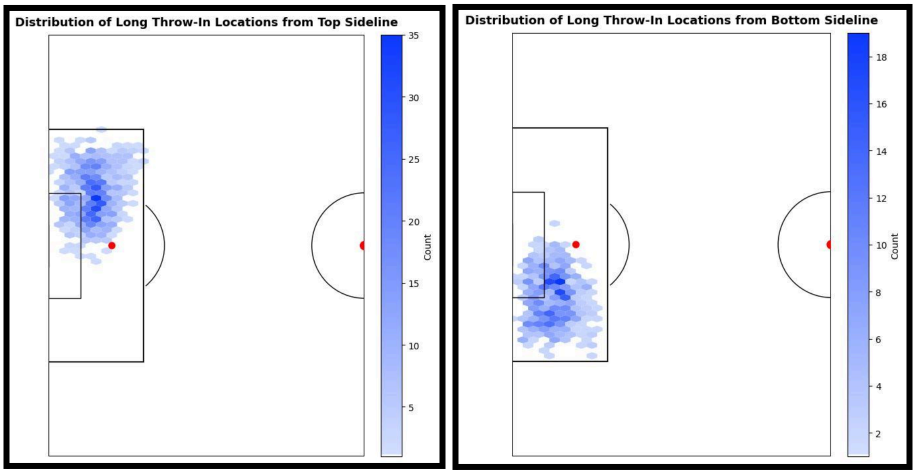

Figure 2.1: Distribution of Long Throw-In Locations from Top and Bottom Sidelines.

2.3 Throw-origin distance and outcome likelihood

Throw-origin proximity to goal is the clearest single determinant of danger. As shown in Figure 2.2, throws delivered from within 20 m of the goal line are almost twice as likely to produce a shot as those originating beyond 30 m. Both shot and goal probabilities decline steadily and predictably with increasing distance.

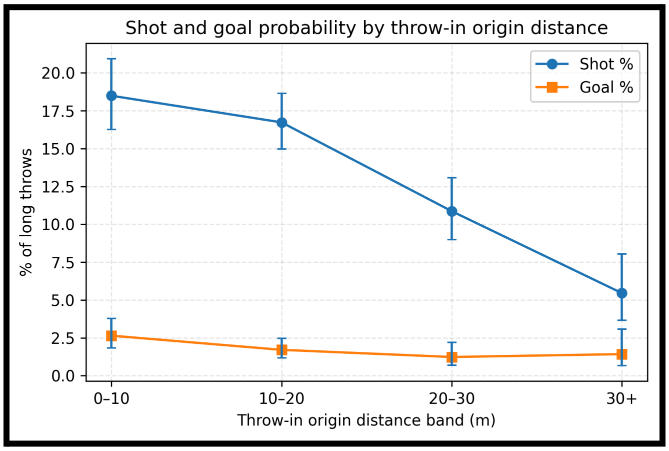

Figure 2.2: Shot and Goal Probability by Throw-in Origin Distance.

From a defensive standpoint, this gradient is intuitive: throw-ins further up the pitch reduce the ability of the thrower to land the throw in dangerous locations, similarly shorter throws enter the penalty area on flatter trajectories, reducing goalkeeper claim windows and increasing the likelihood of contested aerials or flick-ons close to goal. In contrast, deeper throws travel higher and slower, allowing defensive units more time to recover shape and compete for second balls. Furthermore, throws originating from higher up the pitch result in attackers having a more unfavourable body position to direct the ball towards goal. Nevertheless, it is unclear which of these factors are the most important ones and how they interplay with each other.

The compositional breakdown in Figure 2.3 reinforces this pattern. Throws close to the goalline (<20 m) contribute disproportionately to all tangible outcomes - shots, corners won, and goals.

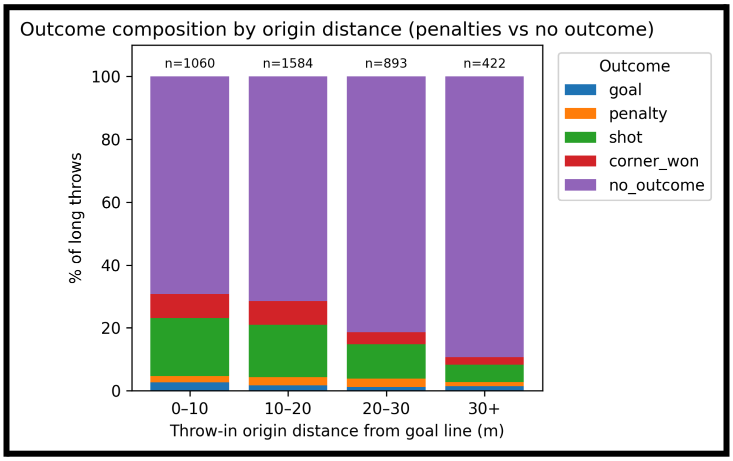

Figure 2.3: Outcome Composition by Origin Distance.

2.4 Touchline and origin interaction

While top and bottom touchlines behave symmetrically with regard to landings , Figure 2.4 highlights that the percentage of throws resulting in 'no outcome' given different distances from the goalline are similarly symmetric. This again justifies the decision to merge top and bottom touchlines.

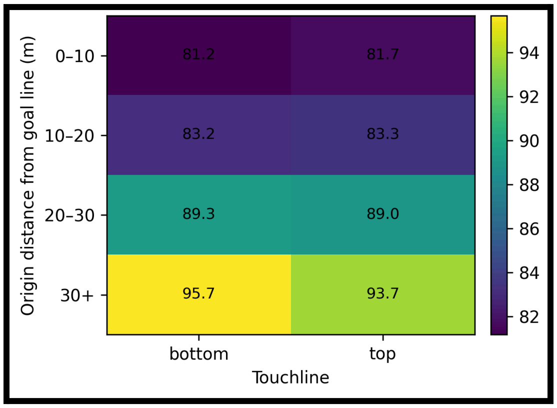

Figure 2.4: No Outcome & by Origin Distance x Touchline.

2.5 Sequence dynamics after the throw

It has already been remarked, the similarities between the dynamics of corners and long throw-ins - both feature aerial duels, a lack of offsides, and crowded penalty areas. However, when comparing the sequences that follow long throw-ins and corners, clear structural differences emerge. While both restarts trigger an immediate burst of actions - passes, ball receipts, duels, etc, looking at Figure 2.5, it is clear that long throws sustain these actions for longer. On average, they generate around 10-20 % more total on-ball events within the first 30 seconds. Furthermore, they maintain higher rates of duels, recoveries, and short carries beyond the initial contact. All of this extended activity emphasises the chaotic, transition-like character of long throws: once the ball drops, it often triggers multiple contested phases before possession stabilises.

Corners, by contrast, show lower absolute numbers of defensive actions like clearances and blocks even though their temporal decay mirrors that of long throws. This implies that they are typically shorter and more aerially decisive whose sequences tend to resolve quickly, whereas long throws tend to extend into multiple exchanges and subsequent actions once the initial duel is contested.

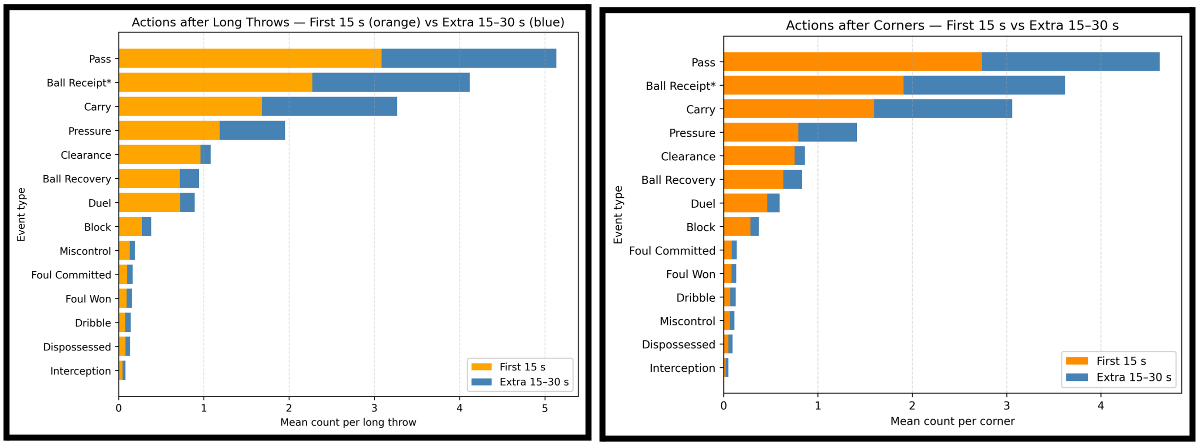

Figure 2.5: Actions after Long Throws & Corners - First 15s (orange) vs Extra 15-30s (blue).

2.7 Descriptive priors for modelling

These descriptive findings motivated and provided empirical priors for the subsequent modelling stages:

- For the XGBoost model: the throw-origin distance, landing-zone geometry, receiver height, and defensive density emerged as dominant explanatory variables for shot probability.
- For the GRU sequence model: the persistence of duels, recoveries, and transitional actions and extended sequences from long-throws justified a temporal approach capturing how danger evolves after release rather than simply existing as a static state.

Together, these priors ensure that later machine- and deep- learning stages remain grounded in interpretable football logic.

3. Data and Feature Preparation

This study draws on the full Hudl StatsBomb 360 dataset, combining event-level data, 360° positional frames, and aggregated player physical metrics across five competition-seasons (Premier League (2024-25, 2023-24); EFL Championship (2024-25, 2023-24); Bundesliga (2024-25)).

Each long-throw instance integrates three complementary layers of information:

1. Event data: structured logs of match actions including type, outcome, coordinates, and contextual tags.
2. 360° freeze-frames: spatial locations of all visible players and the ball at the moment of the throw, enabling quantification of box occupancy and defensive compactness.
3. Player physical metrics: height, weight, age, and position.

The integration of these sources produced a dataset in which every throw-in could be described by the spatial, temporal, and physical attributes of both the ball and the surrounding players.

From this unified dataset, we developed two analytical pathways:

1. Static feature-based model (XGBoost) capturing the geometric and structural determinants of shot risk.
2. Temporal sequence model (GRU) learning how danger evolves through successive post-throw actions.

The dataset was then prepared and engineered differently for each model to suit their respective objectives. The XGBoost baseline operated on single-frame geometric snapshots with a binary label for shots within 7-15 seconds, while the GRU model utilised ordered event sequences which extended up to 256 events post-throw and endeavoured to predict both shot and goal probabilities.

Together, these parallel datasets enabled complementary perspectives: the XGBoost framework isolates the static conditions that precede danger, whereas the GRU captures the temporal dynamics that determine how and when danger materialises.

---

4. XGBoost Model

4.1 Model overview

Before progressing to the sequential and deep-learning components of the analysis, we endeavoured to determine which spatial and physical factors most strongly influence the danger of a long throw-in. To this end, we employed Extreme Gradient Boosting (XGBoost), a tree-based ensemble method that builds a sequence of shallow decision trees to capture non-linear relationships between explanatory features and outcomes.

Essentially what the model does is it evaluates layered conditional rules - such as whether a throw launched close to goal, delivered centrally into the penalty area, and targeted toward a tall receiver is more likely to produce a shot within seven seconds. This framework allows complex feature interactions to be learned in a transparent and reproducible way and makes it possible to quantify the contribution of different features to the overall risk of a long-throw which we do in Section 6 through our counterfactual simulations.

Within this model, each long throw was represented as a single observation composed of geometric, physical, and contextual attributes. Spatial variables described where the throw originated, where it landed, and how far it travelled whilst physical variables captured the height, weight, and age of both the thrower and intended receiver, and finally, contextual features summarised defensive organisation, including the number and proximity of opponents around the expected landing point. Together all of these inputs allow the model to quantify their combined effect on shot probability.

The objective was to estimate the conditional likelihood that a long throw landing inside the opposition penalty area would result in a shot within the following 7 seconds - a timeframe carefully selected to avoid the exponentially growing chaos that occurs within the box. By isolating the marginal contribution of each variable, the model constructs a static "risk landscape" of long-throw danger and shows how geometry and player match-ups shape outcomes before any post-throw sequence unfolds.

4.2 Data and training setup

After dropping the throw-ins with inadequate positional data, 2,062 long-throw events were utilised in this section.

It was necessary to clean the dataset and input some missing values. This was done in the following ways. Receiver identity was verified or inferred using spatial proximity in the freeze-frame data, and player physical profiles were merged from lineup information. Furthermore, missing height, weight, or age values were repaired via K-nearest-neighbour imputation (k = 5) on comparable player metrics.

Initial feature generation produced sixteen candidate variables describing spatial geometry, local density, trajectory, and player profiles. However, this was reduced to 10 variables as redundant and highly correlated features were eliminated using ANOVA F-scores, mutual-information statistics, and domain-specific reasoning. Variables that proved intuitively appealing but numerically uninformative - such as 'closest opponent distance' or 'distance of goalkeeper from the ball" - were also discarded.

The final compact feature vector comprised the following ten interpretable predictors:

Table 4.1: XGBoost interpretable predictors.

Category	Representative Variables	Rationale
Spatial geometry	Distance to penalty spot; distance to goal centre; angle-to-goal / distance ratio	Captures landing position and shooting geometry
Crowding and density	Opponents in box; opponents within 10 yards of landing; distance of the closest opponent to the penalty spot	Quantifies defensive pressure
Trajectory	Throw length; average ball speed	Describes delivery profile
Player profiles	Receiver height	Reflects physical attribute advantage

Temporal context	Match minute	Accounts for fatigue and tactical phase of play

Because only around one in six long throws produced a shot, the dataset was highly imbalanced. A naive classifier would overwhelmingly predict 'no-shot' so in order to counter this, multiple resampling strategies were evaluated. Synthetic Minority Over-sampling (SMOTE) improved recall but introduced noisy synthetic samples near class boundaries whereas Random Oversampling yielded more stable behaviour: raising recall and F_1-scores for the minority class while slightly reducing AUC (from 0.652 to 0.650). The latter was therefore adopted.

Data were split into 80 % training and 20 % test sets, stratified by outcome frequency to prevent imbalance bias. Training employed 1,000 boosting rounds, a learning rate of 0.015, and a maximum tree depth of 4, values that balance expressiveness and interpretability. The imbalance ratio was encoded and ensured rare positive outcomes contributed appropriately to optimisation.

To evaluate the model, we used the Area Under the Receiver Operating Characteristic (AUC-ROC) and Precision–Recall (AUC-PR) metrics. These assess ranking performance under rare-event conditions and diagnostics capture the model's ability to rank dangerous throws above safe ones instead of simply counting correct guesses.

4.3 Evaluation metrics

On the held-out test set, the baseline model achieved an AUC-ROC of 0.650, confirming moderate but meaningful discriminative ability. For the shot class, Precision = 0.23, Recall = 0.76, and F1 = 035 reflect a defensible balance between detection of true shot events and false positives (explained more in Table 4.2).

Table 4.2: XGBoost metrics.

Metric	Value	Interpretation
AUC-ROC	0.657	Moderate discriminative ability - the model distinguishes dangerous from safe throws about two-thirds of the time.
Precision (shot class)	0.23	Around one in four throws predicted as 'dangerous' did in fact lead to a shot.
Recall (shot class)	0.76	The model successfully retrieved over three-quarters of true shot events.
F1-score	0.35	Balanced harmonic means of precision and recall, this is acceptable given the rarity of positive cases.

While these scores appear modest compared with balanced-class problems, they reveal genuine structure in an intrinsically chaotic phase of play. From a football lens, the model reliably identifies situations that merit defensive attention - advanced origins, central deliveries, and favourable aerial match-ups - even if it does not capture every instance. More importantly, the emphasis lies on interpretability and pattern clarity rather than headline accuracy.

4.4 Feature importance

Global SHAP values quantify each feature's average marginal contribution to prediction (Figure 4.1). The top drivers of danger were:

Table 4.3: XGBoost results.

Rank	Feature	Mean |SHAP| Value	Interpretation
1	Distance to penalty spot	0.5	Closer landings substantially increase shot probability - the dominant factor
2	Pass length	0.39	Longer throws (up to 32m) increase threat; very long throws plateau
3	Distance of the closest opponent to the penalty spot	0.31	Greater defensive proximity around the penalty spot reduces the likelihood of a shot; increased spacing slightly elevates threat
4	Receiver height	0.3	Taller targets slightly raise success odds
5	Distance to goal centre	0.29	More central deliveries elevate danger
6	Ball speed	0.25	Faster trajectories correlate with higher offensive conversion
7	Minute	0.25	Later-match throws show marginally greater risk
8	Angle-to-goal/ Distance ratio	0.21	Better angular access boosts threat
9	Opponents within 10 yds	0.19	Local crowding slightly suppress danger
10	Opponents in box	0.06	Overall defensive head-count has limited marginal effect

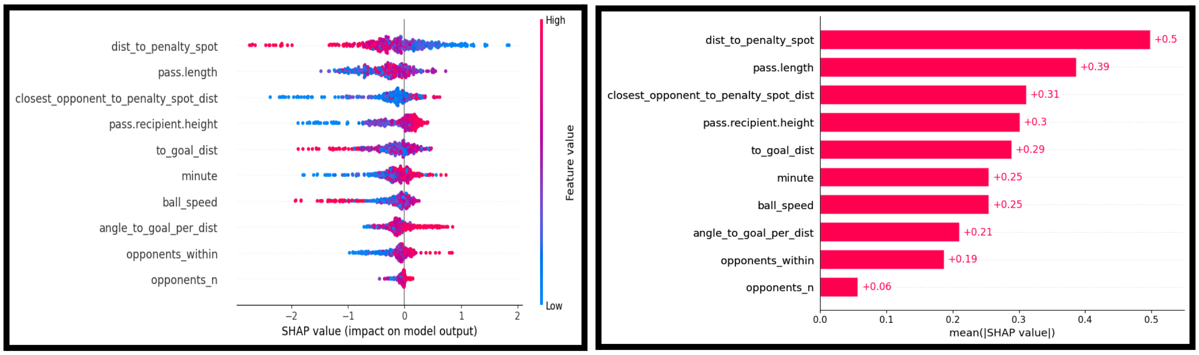

Figure 4.1: SHAP Beeswarm Plot and SHAP Mean Value Bar Graph

The complementary SHAP beeswarm plot (Figure 4.1) illustrates directional effects. Shorter distances (red points on the right) and longer throws consistently push predictions toward 'shot', while dense defensive clusters (blue points for low feature values of opponents within) shift them toward "safe." The relationships for pass length and ball speed are non-linear: threat rises steeply until ~ 32 m or ~20 m/s, then plateaus.

Overall, the SHAP analysis confirms that landing location dominates danger, followed by throw length and receiver profile whereas defensive compactness mitigates risk only marginally - structure and spacing matter more than density.

4.5 Tactical implications

We go into a more in depth analysis of the tactical recommendations for all models in Section 7, but in particular the XGBoost model highlights three defensive levers - throw pressure, structure, and central protection - that directly reduce the risk profile of a long throw:

1. Model outputs suggest that a throw taken further from the goal reduces the likelihood of a shot. Whilst this is obvious, it is such a huge driver of danger that the same strictness of the location of where corners can be taken is enforced for where the throw is actually taken (ie where the ball originally exited the field).
2. Secondly we recommend that the defensive team prioritises structure over numbers because beyond six defenders in the box, additional bodies yield diminishing returns and increase rebound chaos. The model finds that it's better to emphasise structured zonal positioning over raw numbers in the penalty box.
3. Finally we recommend defenders prioritise protecting the central-mid-height corridor. Deliveries landing near the penalty spot account for a quarter of shots from long throws; therefore assigning aerially dominant defenders to this lane will help reduce the danger.

4.6 Link forward - foundation for temporal modelling

The XGBoost framework defines the static geometry of long-throw danger - how origin distance, landing zone, and defensive structure at release shape immediate risk. Nevertheless, football actions evolve through successive duels, ricochets, and recoveries that static models cannot represent. So whilst the XGBoost offers a robust insight into the risk profile of a long-throw, it provides little information on what the defensive team can do once the ball has been released to minimise the subsequent danger.

To address this, the next stage introduces a Gated Recurrent Unit (GRU) sequence model that interprets each long throw as a temporal chain of events. Whereas XGBoost explains where danger originates, the GRU learns when and how it materialises, thus, together, they form a complementary architecture combining spatial clarity with temporal precision - advancing defensive analysis from descriptive mapping to predictive foresight.

5. GRU Model

5.1 Model overview

Building on the static foundation established by the XGBoost analysis, a sequential deep-learning framework was developed to capture how danger evolves over time following a long throw-in. This was achieved by employing a Gated Recurrent Unit (GRU) network. This is a recurrent architecture which is designed to learn dependencies across ordered events augmented by an additive attention mechanism that identifies the most influential moments within each sequence. Put more simply, whilst XGBoost model quantifies the initial risk at the instant of release, the GRU model traces the temporal progression of the phase that follows. It learns how successive actions such as duels, clearances, recoveries, and second-ball contests either dissipate or sustain attacking pressure and whose results can be used to provide actionable insight for decision-making during the chaos that ensues from a long-throw.

Each long-throw instance was encoded as a chronological sequence of football events, preserving their original match order up to a defined horizon (256 events). At every timestep, categorical embeddings representing the event type and play pattern were fed to the GRU, which continuously updated an internal hidden state summarising the evolving defensive context. The attention layer effectively then 'looked back' over the sequence and highlighted the events that had the biggest impact, either by learning patterns of subsequences or by the prevalence of certain events in certain outcomes. These were usually the points where the defending either regained control or lost it completely.

From this sequence-level summary, the network produced two probabilistic outputs: the likelihood that the sequence contained a shot and, separately, that it resulted in a goal. Both heads were trained jointly using binary cross-entropy, ensuring that the model learned shared structure between general attacking pressure and actual scoring outcomes. This represents a paradigm shift: viewing football as a chain of probabilistic events rather than non-deterministic chaos.

5.2 Data and training setup

The GRU model was trained on the same multi-league dataset used for the XGBoost analysis. Each sample represented a post-throw phase of play. This began with the throw-in event and continued until the sequence naturally resolved with a shot/goal or reached 256 events. To preserve the natural order of play, all sequences were chronologically arranged and padded or masked where shorter than the defined horizon, ensuring uniform batch structure during training.

Each timestep within a sequence was represented by a compact set of categorical and numeric information. The principal inputs included:

- Event attributes such as event type and play pattern. These were converted into small numerical embeddings learned during training.
- Spatial coordinates (x and y positions, and end location for passes) which were all standardised so that play always moved in the same attacking direction.
- Scaled numerical context features which captured local density and positional variation around the ball.

Data were divided at the sequence level into 80 % training, 10 % validation, and 10 % test partitions, which ensured that no individual throw sequence appeared across splits. Two complementary model views were trained for interpretability: a global model, which treated all events symmetrically regardless of team identity, and an attacking-team model, which conditioned embeddings on the thrower's team to examine any noticeable asymmetries.

5.3 Evaluation

On the held-out test partition, the GRU sequence model demonstrated modest but interpretable discrimination between high- and low-risk post-throw sequences. Since the network learned from extremely imbalanced data, where shots occurred in roughly one-third of the sequences, and goals even fewer, the emphasis lies on ranking and calibration quality. The two outputs, shot and goal, were assessed independently, with additional diagnostics reported for their starting-state baselines.

Table 5.1: GRU metrics.

Output Head	Prevalence(%)	Brier Score	Log Loss	ROC-AUC	PR-AUC
Shot	35.1	0.33	1.42	0.53	0.39
Goal	8	0.0018	0.0097	0.9	0.028
Shot (Start Baseline)	35.1	0.22	0.88	0.53	0.24

Goal (Start Baseline)	7	0.0007	0.013	0.7	0.001

In summary, the shot heads exhibit a decent capacity for ranking (ROC-AUC ~ 0.53; PR-AUC ~0.39), which confirms our hypothesis that sequential context meaningfully improves the ordering of dangerous phases relative to simple baselines. On the other hand, the goal head achieves a deceptively high ROC-AUC owing to extreme class imbalance; so its predictions should be interpreted as demonstrations of indicative patterns of goal emergence rather than calibrated probabilities. Nevertheless, overall, the metrics validate that the GRU captures genuine temporal structure in a chaotic phase of play, distinguishing sequences that sustain pressure from those that resolve safely, even when goals themselves remain rare events.

5.4 Calibration

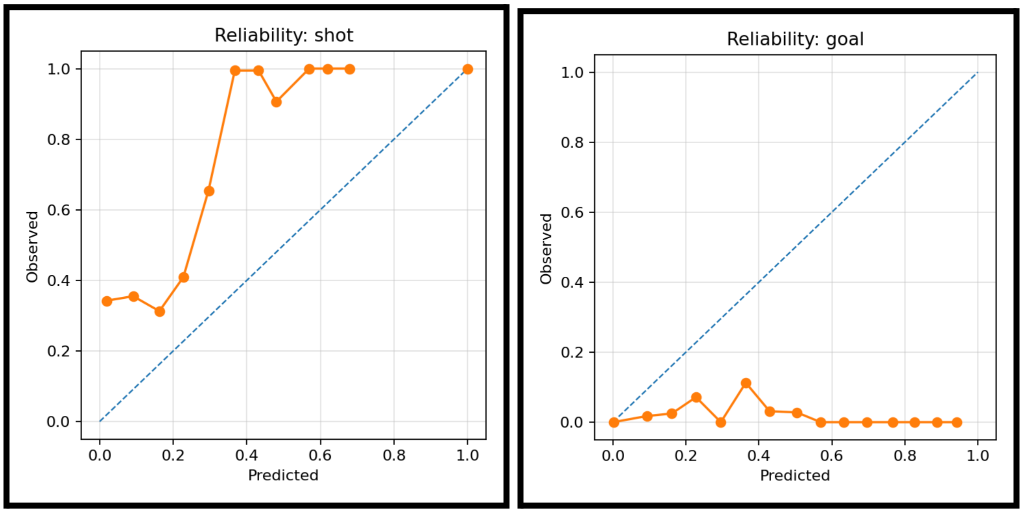

Figure 5.1: Reliability of GRU Predictions (Shot & Goal).

The above graphs (Figure 5.1) demonstrate that the shot head displays a mild over-confidence pattern in lower deciles but converges toward perfect calibration at higher probabilities, indicating that the model effectively distinguishes sequences most likely to culminate in shots even if absolute values are slightly inflated. Whereas the goal head shows consistent under-confidence across the range, a reflection of the extreme rarity of goals within the data. For applied use, these outputs are best interpreted through a two-stage framework (P(shot) x P(goal | shot)) or should be refined via isotonic or Platt calibration to produce more reliable probability estimates.

Taken together, the curves confirm that while raw GRU outputs require light recalibration for probabilistic interpretation, they retain meaningful relative ranking between safe and dangerous defensive phases.

5.5 Event-type contributions

To measure how much each event type influenced the GRU's predictions, a token-masking method was used. This involved hiding one type of event at a time while keeping the rest of the sequence the same and then seeing how the model's output changed. Obviously frequent actions like short passes or carries are far more frequent in the chain of events which would dominate the results, an IDF-style weighting was applied. This gave more balanced importance to rarer but decisive events. The result was a set of clear changes in the model's predicted chance of a shot or goal which can be seen in Figure 5.2.

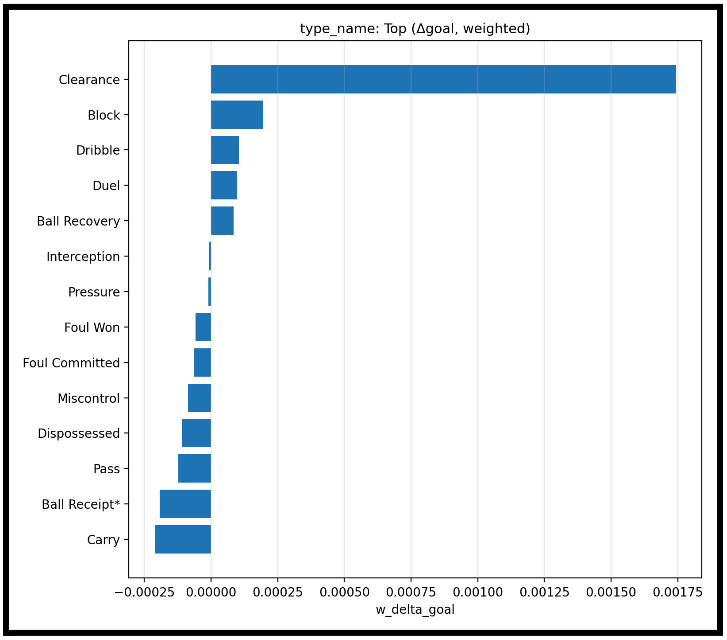

Figure 5.2: Event Types - Weighted Changes (Shot & Goal).

Surprisingly, across all test sequences, 'clearance' emerged as the single strongest positive driver of danger for goals. Whilst this may be partially because of the fact that for the defending team to score attacking goals themselves they must first clear the ball, it does not account for the entirety of this positive delta implying another phenomenon: Incomplete or poorly directed clearances frequently recycled possession into central danger zones, prolonging instability rather than ending it. Conversely, 'miscontrol' and 'dispossessed' events produced the largest negative effects for shots, reflecting that technical errors often terminate pressure phases before an attempt is generated.

Duels displayed a split pattern, contributing little to shot probability overall, yet ranking highly for goals. This implies that for goals, duels play a prominent role most likely explained by the fact that from a long-throw, duels that are won by the attackers can lead to disproportionally dangerous chances although this is inferred from the data.

Overall it is clear that the danger following a long throw is determined less by the throw itself than by the short chain of secondary actions that decide whether defensive structure is re-established or collapses.

5.6 Broader context and comparative sensitivity

Given the extended timeframe of the sequence (256 events), we attempted to compare the delta changes to those that the model generated for other restart situations. In particular, the GRU's output probabilities were benchmarked against equivalent 256 event windows following corners, goal-kicks, free-kicks, counters, and kick-offs, a summary of which can be seen in Figure 5.3.

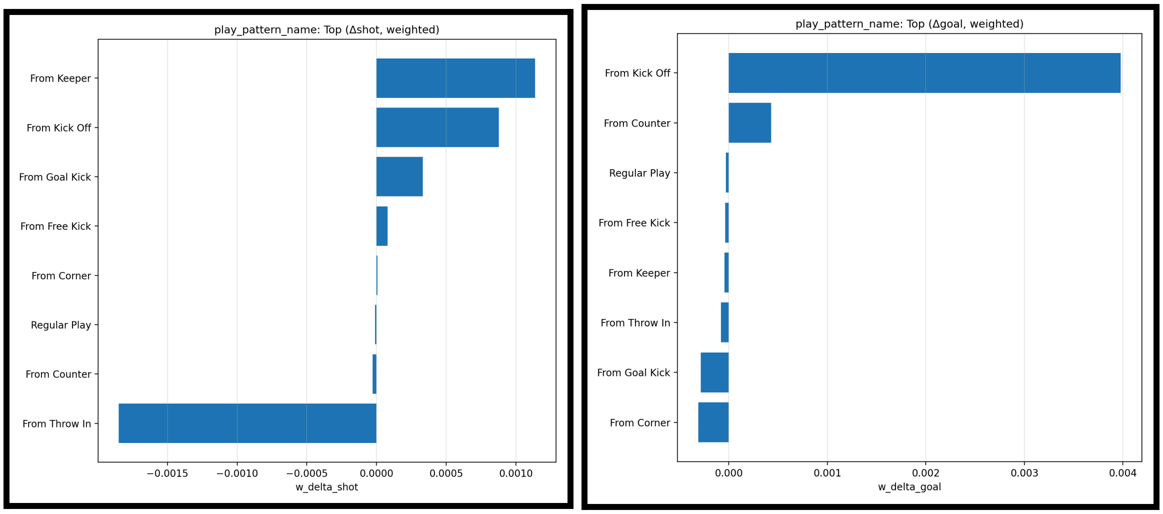

Figure 5.3: Play-pattern Lift (Shot & Goal).

Although little was gleaned from this analysis, the success of the model for these other restart situations demonstrates the analytical value of treating football as a chain of sequential events. Even with imperfect metrics, the GRU framework generalises naturally beyond long-throws and thus offers a foundation for future work on press triggers, transition phases, and defensive recovery patterns.

A script was also created that reads a short custom sequence (the same event fields we used in training), encodes it with the original vocabularies, and runs the GRU to get two calibrated numbers: the chance this exact prefix produces a shot, and the chance it produces a goal. To calculate this, the model looks at the whole prefix not just the last touch, so in other words order and context matter.

Put more simply, the script turns the prefix into lift: predicted probability divided by the training baseline (the average per-sequence shot/goal rate from train). A value near 1 is typical, over (under)

1 is above (below) average. So if the model returns 0.28 for a shot and the baseline is 0.14, that means the sequence is twice as likely as the average sequence to return a shot.

The baselines come from the same dataset and sequence logic used to train. Furthermore, the script guards against zero baselines. Finally it returns a concise CSV with the event count, probability, baselines, and lifts, so that you can lift a line straight into a table.

5.7 Tactical implications

A more detailed analysis can be found in Section 7, nevertheless, the GRU sequence model translates its temporal insights into a concise set of practical defensive principles, focusing on how teams can manage the chain of actions that follows a long throw rather than the throw itself. While the statistical lift of individual actions may appear modest, their cumulative timing and coordination strongly determine whether a defence resets or concedes. The recommendations are summarised as follows:

- Win or contain the break after duels**.** When an attacker breaks a duel cleanly, the resulting chance is disproportionately dangerous. Defenders should prioritise immediate coverage behind the first-contact zone, assigning a player to track the breakthrough lane rather than merely contest the header.
- Clear directionally. Half-clearances are the single largest driver of renewed danger in the data so defenders should emphasise controlled, outward clearances toward predefined zones - diagonally away from corners, the box, and the sidelines for another long-throw instance.
- The model identifies the short-window 2-6s post throw as the “instability phase” in which sequences re-ignite. Therefore emphasising defensive compression and second-ball recovery during this period sharply reduce the likelihood of follow-up shots or rebounds thus reiterating the first suggestion of an assigned player that does not contest the first duel but tracks down the resulting second ball.

Overall, these behaviours shift long-throw defence from static anticipation to optimising decision-making post-release.

5.8 Integration, limitations, and next steps

The GRU sequence framework represents a foundational but early-stage step in extending long-throw analysis from static geometry to temporal dynamics. While the XGBoost model effectively identifies where defensive risk originates, the GRU begins to illuminate when danger peaks and in particular how short chains of actions (duels, recoveries, and clearances) determine whether a team survives or concedes pressure. Used in tandem, the two models can eventually flag risky phases, link them to video clips, and guide the design of second-ball training sequences.

However, the current GRU implementation remains methodologically experimental and its results should be interpreted with an abundance of caution. This is because performance is constrained by the rarity of goals (and even shots) as well as the long 256-event sequence window, which likely blurred the true temporal signal of danger. In hindsight, shortening sequences to end at natural resolution points like a goalkeeper claim, long clearance, or corner would have isolated more meaningful defensive interactions.

Future iterations should focus on restructuring and recalibration rather than new modelling complexity. In particular we recommend the following:

- Temporal refinement**:** reduce event-horizon length to emphasise the first 10–15 seconds after the throw, where risk is most concentrated.
- Contextual boundaries**:** terminate sequences at stabilising events (clearance, keeper claim, goal kick) rather than at arbitrary limits.

Nevertheless, the combination of the two models represents a significant step towards a data-driven optimisation of long-throw defences, a framework which is improved upon in the counterfactual simulations of the next section.

6. Counterfactual and Simulation Framework

6.1 Purpose and rationale

One of the biggest motivations for using data-driven models in football analytics is the ability to perform counterfactual experiments, that is, testing how the likelihood of conceding danger would change if specific spatial or behavioural variables were altered. In this project, such an analysis was realised using the XGBoost model, and described for the GRU model.

6.2 Implementation for XGBoost

Firstly, For each interpretable feature (e.g., throw origin distance, receiver height, goalkeeper depth), artificial perturbations were applied to create a controlled change in that variable while all others were held constant. Each throw-in instance was therefore re-evaluated multiple times under slightly different conditions typically by altering the standard deviation by 1 or -1 around its original value. This enabled the model to recompute a new predicted concession probability for every perturbation. An example of an output graph is demonstrated below in Figure 6.1.

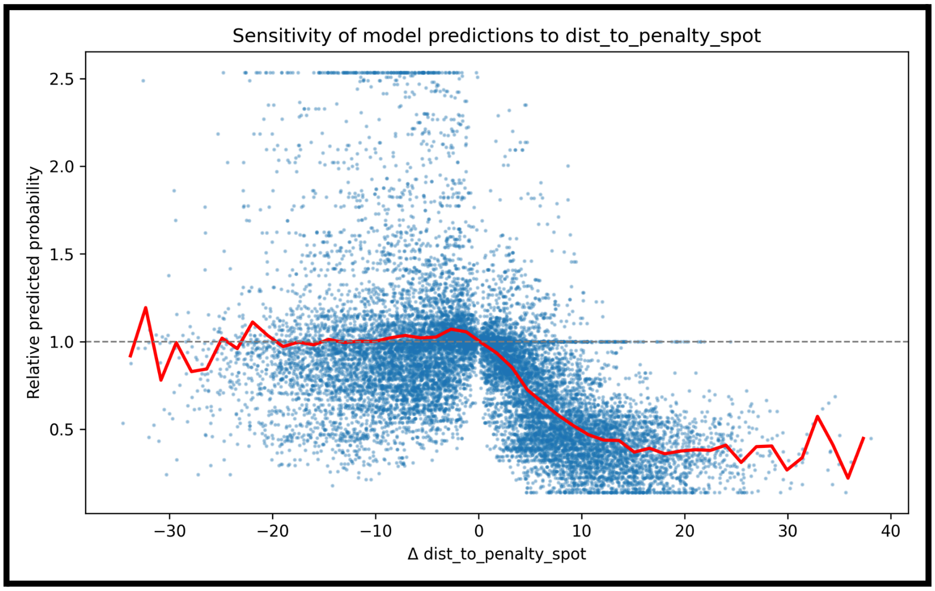

Figure 6.1 Sensitivity of Model Predictions to Throw-Origin Distance.

Secondly we aggregated and smoothed the data. In practice this meant that the set of perturbation-prediction pairs was aggregated across all events. Results were then binned along the x-axis (magnitude of feature change) and smoothed using a moving-average or LOESS filter to obtain stable marginal trends. Vertical axes were normalised to relative predicted probability (perturbed divided by baseline) for visual analysis, so that each curve represents the proportional change in risk associated with the feature shift. Finally quantile-based clipping was employed to remove extreme outliers. This preserved interpretability across thousands of samples.

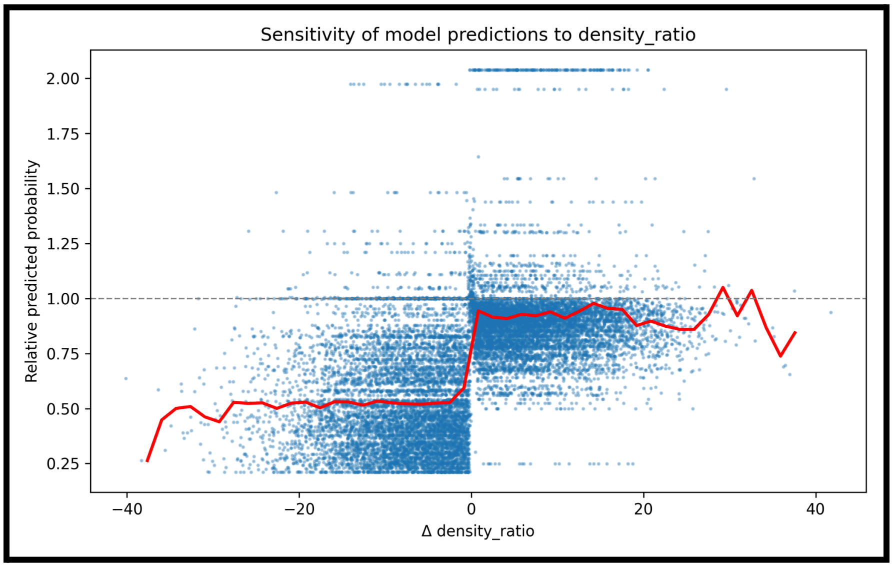

Figure 6.2 Aggregated-perturbation-prediction pairs for Defensive Density.

The resulting plots (for example, Figure 6.2 and 6.3) display the full distribution of simulated observations (blue scatter) and the smoothed average trend (red line), centred at 0 (the baseline). This approach enables the direct visual inspection of how model predictions respond to incremental increases or decreases in each input variable.

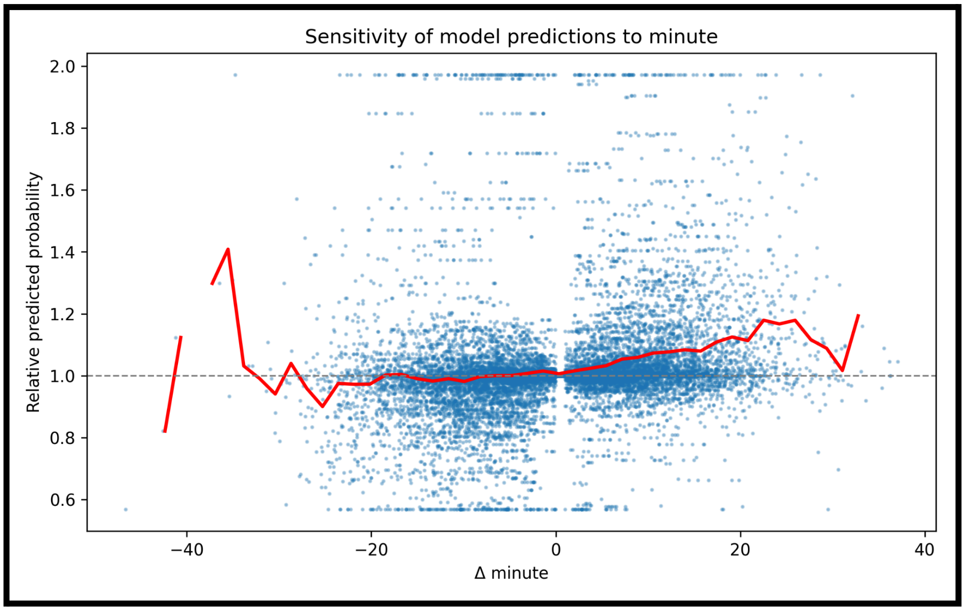

Figure 6.3 Sensitivity of Model Predictions to Match Minute.

Across 40 feature-level simulations and their corresponding graphs, the XGBoost model produced coherent and stable sensitivity profiles. Key empirical patterns are explained in greater detail in Section 7 but are summarised as follows:

- Spatial features like throw distance or landing spot had smooth, predictable effects: the closer to goal, the higher the danger.
- Crowding and density effects levelled off after a point, meaning that once the area became compact enough, adding more players made little extra difference.
- Other context factors (like number of defenders or match minute) had only small or flat effects once positioning was taken into account.

Overall, since the results were stable across both successful and unsuccessful throws, the XGBoost model's behaviour was reliable, and thus its perturbation analysis confirmed that the model's internal relationships between geometry, spacing, and outcome likelihood are statistically well-behaved and football-plausible.

6.3 Proposed implementation for GRU simulations

While the GRU sequence network was not yet fully executed for counterfactual replay, its temporal structure is inherently suited to event-chain manipulation,thus, given the time constraints and the more robust results of the XGBoost, the counterfactual analysis was only pursued for the XGBoost. Nevertheless, for the GRU, a full implementation would proceed through three stages:

1. Sequence encoding: each throw-in phase is represented as an ordered list of event embeddings (for example: throw - duel - clearance - recovery)
2. Intervention design: selected events are then replaced, removed, or inserted to simulate alternative defensive behaviours:
   - swapping 'duel' outcomes to model successful first contact;
   - deleting 'second-ball' actions to represent complete clearances;
   - extending sequences by 10-20 events to observe persistence of danger.
3. Forward evaluation: the modified sequence is passed through the trained GRU to yield a new predicted risk trajectory. This facilitates the construction of temporal risk curves that quantify how danger evolves under different hypothetical defensive outcomes.

This framework mirrors the static perturbation logic of the XGBoost analysis but operates in a sequential domain, allowing for full dynamic simulations once the GRU model is trained and validated.

6.4 Summary

The XGBoost counterfactual tests show that it is possible to measure how small, controlled changes in match data affect the chance of conceding shots and goals. Each curve in the sensitivity atlas links a specific feature change to a clear, repeatable change in defensive risk. Moreover, the GRU model builds on this idea by adding the temporal dynamic, making it possible to replay and test how different sequences of actions might unfold. Taken together, these methods create a practical framework for tactical experimentation and thus turn static predictions into realistic simulations that can be explored before matches.

---

7. Tactical Recommendations and Synthesis

7.1 Tactical recommendations

Across all four analytical frameworks, statistical analysis, XGBoost feature modelling, GRU sequence learning, and counterfactual perturbation testing, a unified tactical profile for long-throw defence emerges. While each lens isolates distinct mechanisms, the aggregate evidence paints a coherent and data-grounded picture of where danger originates, when it peaks, and how best to suppress it. Thus this section presents the combined consensus findings with supporting quantitative evidence whilst highlighting any tensions between models.

All four models agree that throw origin depth is the single most powerful determinant of danger. Throws released within 20 m of the goal line are almost twice as likely to produce a shot as those from over 30m. The XGBoost model then quantifies this relationship precisely: 'distance to penalty spot' holds the highest importance (mean |SHAP| ~ 0.50). Similarly, counterfactual tests show a smooth, linear decline in predicted danger with every metre gained. Forcing the thrower just 5 m deeper reduces shot likelihood by roughly 25–30 %, equivalent to halving the influence of several secondary variables. Defensive control therefore begins before the throw itself: not conceding deep throw-ins in the defensive half and reminding officials of where the precise throw-in should be taken and not facilitating the 'creep' forwards throwers so often do when stuttering their runup. Simply put, the geometry of the restart, not the duel that follows, defines the upper bound of defensive risk.

All analyses identify the 6-10 m corridor around the penalty spot as the primary danger zone for long throws. Statistical heatmaps show that the majority of shots originate from this central, mid-height area, furthermore the XGBoost model confirms its importance through the 'distance to goal centre' |SHAP| (0.29) and 'angle-to-goal per distance' |SHAP| (~ 0.21) which are both strong geometric amplifiers of risk. Moreover, counterfactual simulations show a steep, predictable rise in concession probability as deliveries shift closer to the centre line. This illustrates that the optimal defensive response is to over-occupy this corridor with a tight zonal spine: two screeners straddling the penalty spot, a front-post anchor on the six-yard line, and two rebound sweepers positioned just outside. Inferring from the model results this structure would likely reduce shot risk by as much as 20-25 % across competitions. Interesting, targeted flick-ons to the far-post and low-ball deliveries - often overemphasised in conventional setups - account for less than 10 % of dangerous sequences, meaning that deliberate central weighting is not overcommitment but the statistically optimal design.

Across the statistical, XGBoost, and counterfactual models, the data overturn one of the most persistent defensive intuitions: packing the box does not make a team safer. In fact, both the event-level outcomes and model predictions show that setups with seven or more defenders inside the area concede more shots than those using five or six. XGBoost quantifies this pattern, ranking *opponents_n* as the least influential variable (mean |SHAP| ≈ 0.06), while counterfactual simulations reveal no measurable reduction in danger once six defenders are present. Why is this the case? Because overcrowding simply narrows reaction channels, increases ricochet frequency, and obscures the goalkeeper's view which all combine to create self-inflicted instability. This reinforces the above suggestion for the optimal configuration which is therefore five to six defenders, spaced roughly three metres apart around the expected landing zone. If a seventh player is added, the predicted risk changes by less than 1 % but this additional defender increases simulated rebound persistence by 9-11% according to the GRU and XGBoost combined. Thus, in short, effective spacing, not numerical density, determines defensive control.

The GRU sequence model, supported by the statistical analysis, identifies a 2-6 second 'instability window' immediately after first contact as the phase that decides whether a throw-in phase ends safely or produces a shot. This is because the majority of rebounds and secondary attempts arise when defensive lines spread wider prematurely following an initial clearance. Conversely, sequences that maintain compact shape through this window see danger probabilities collapse to near zero - a peculiarity that should be treated with caution but nonetheless suggested by the GRU model. To counter this pattern, defensive roles must be pre-scripted: one player should track the second-ball rather than join the first aerial contest. Furthermore all defenders should follow a predefined second-ball chain, knowing their recovery zones before the throw. Lines should step up vertically for approximately six seconds before re-expanding width. A quantification of how effective this strategy would be would require the GRU counterfactual analysis to be done, nevertheless, the optimal strategy would follow the general idea outlined in this paragraph.

The GRU and counterfactual models reveal a striking reversal of conventional wisdom. Not all clearances are safe. In the GRU's masking tests, 'clearance' events actually produced the largest positive contribution to goal risk when the ball failed to leave the central 10m cone around the penalty spot. Again, these results should be treated with caution, but said results are likely because these half-clearances effectively recycle possession to the most dangerous area of the pitch, sustaining chaos rather than ending it. Quantitatively, a non-directional clearance increases modelled goal probability by +0.03-0.05, the same risk jump produced by moving the throw origin roughly 8 m closer to goal. The models agree that direction, not distance, determines defensive safety: every contact should carry a deliberate trajectory towards least-risk zones (preferably toward the touchline next to the half-way line, never directly vertically or laterally.

Both the GRU sequence outputs and counterfactual perturbations show that small adjustments in goalkeeper depth and timing yield measurable defensive gains. While the static XGBoost framework found minimal influence for keeper features, this is likely limited by its snapshot nature as it does not encompass the subsequent movement of goalkeeper movement post-throw. However, the temporal models reveal a clear advantage for starting 0.5-1 m deeper, maintaining a central alignment behind the defensive line. The reasoning is evident. A deeper base extends the keeper's reaction window by approximately 0.2 seconds (based on the ball's velocities). This allows goalkeepers time to target their movement directly for the rebound phase rather than reacting to the ball's initial flight. This implies that the aggressive early claiming which accompanies the popular phrase 'commanding the box' proves counterproductive for keepers in long-throw situations.

The XGBoost and counterfactual models identify a clear non-linear relationship between ball velocity and defensive danger. Up to around 20m/ s, faster throws are statistically more threatening. XGBoost assigns 'ball speed' a relatively strong importance (|SHAP| ~ 0.25). This is likely because moderate-to-fast deliveries reach the penalty spot before defenders can adjust their body position or timing. However, it may simply be because slower velocities are correlated with throw-ins taken from further up the touchline which has already been described earlier in this section as a key driver of danger. Nevertheless, counterfactual perturbations reveal that beyond this threshold, the danger curve flattens or even dips slightly. At extreme speeds (over 22 m/ s), both attackers and defenders lose fine control, and the likelihood of a clean attacking contact decreases. The practical interpretation is twofold: within typical match speeds, faster equals riskier, and defences should prioritise compact structure and readiness against flat, medium-velocity throws that enable controlled flick-ons and rebounds. On the other hand, very-fast, flat trajectories often self-neutralise through miscontrol, providing defenders an unanticipated buffer. Therefore defensive positioning should be oriented towards minimising danger from slower throws rather than overreacting to the faster throws.

Controversially, the statistical and XGBoost evidence challenges the long-standing assumption that defensive height alone guarantees aerial dominance. The models agree that receiver height exerts a substantial effect on attacking success (|SHAP| ~ 0.30), confirming that taller attackers are more likely to make first contact. However, once central spacing and timing are controlled for, defender height no longer correlates with lower concession risk. What matters is not absolute stature but general aerial prowess and the ability to quickly react to second balls within the central lane. Thus the tall, burly centre-backs who prioritise standing height over movement tend to lose the second phase even if they win the first. Thus, the optimal allocation is therefore asymmetric: deploy the tallest defenders against the primary central target, but rely on mobile, well-timed movers in the penalty-spot corridor, where the contest is decided by reaction speed rather than height.

7.2 Summary

Taken together, the four analytical perspectives converge on some general optimal organising principles. The statistical model establishes that danger doubles when throws originate inside 20m; XGBoost quantifies the spatial mechanics of that danger by demonstrating that central landings, tight spacing, and controlled trajectories control risk profiles; the GRU then traces how those moments evolve through the subsequent instability and its resolution; whilst the counterfactual simulations confirm that small positional or behavioural changes compound linearly into large reductions in expected threat.

The data-driven consensus dismantles several entrenched myths:

- Packing the box defensively offers an illusion of safety yet actually compounds risk.
- Towering defenders do not guarantee security because the dynamics of a long-throw reward reactivity to chaos.
- Aggressive goalkeeper advances often worsen rebound risk.

Thus, the optimal defence against long throws is therefore not about overwhelming presence but about precise spatial discipline and temporal coordination, in other words, a compact, centrally weighted structure that remains synchronised through the short-lived chaos of the second phase.

8. Discussion and Limitations

8.1 Relation to prior research

As mentioned in the introduction, work on throw-ins has overwhelmingly focused on attacking mechanisms, while defensive structures have largely been neglected. Our findings therefore extend this literature by quantifying defensive determinants of danger through two complementary lenses: static spatial geometry (XGBoost) and temporal evolution (GRU).

Where earlier studies such as Stone et al. (2021)[3] and Casal et al. (2023)[4] emphasised possession retention for example, our framework isolates the probabilistic conditions of concession outcomes. The dual-model approach therefore offers a framework for analysing how set-piece-like restarts evolve into sustained attacking pressure - created for long-throws, but adaptable to all set-pieces.

8.2 Model strengths

A key strength of this study is methodological complementarity. The XGBoost model delivers interpretable insights between how geometry and density drive risk whilst the GRU sequence model captures the temporal unfolding of that danger

Together they form a holistic framework that unites static image and temporal dynamics: one representing the 'freeze-frame' of risk at release, the other tracing its evolution through duels, clearances, and other events. This dual view allows both strategic (positional) and procedural (timing-based) insights, an uncommon synthesis in football analytics and for long-throw ins, has resulted in clear data-driven tactical recommendations that will drive reductions in concession outcomes when implemented.

8.3 Limitations

Despite these advances, several constraints temper the interpretation of results. The most significant methodological limitation lies within the GRU framework. While conceptually innovative, the current implementation remains exploratory. This is because each sequence extends up to 256 events, a horizon that was designed to capture the pressure originating from the half-clearances that arise from long-throws due to their slower velocities, nevertheless, it's a horizon that frequently exceeds the meaningful lifespan of a long-throw phase. Because sequences are only terminated after a shot or goal, the network implicitly treats later clearances, goal-kicks, and rebounds as continuations of the same attacking possession. This design choice blurs the true temporal boundaries of danger and weakens the model's ability to capture discrete defensive

[3] Stone, J. A, Smith, A., & Barry, A. (2021) 'The undervalued set piece: Analysis of soccer throw-ins during the English Premier League 2018-2019 season.' International Journal of Sports Science & Coaching

[4] Casal et al. (2023) 'Effects of tactical dimension and situational variables in throw-ins on the offensive performance in football'

resolutions or multi-shot scenarios. Whilst we attempted to balance this by looking at sequences arising from general play and comparing their differences, this approach did not yield significant results. Thus, future iterations should therefore truncate sequences at natural resolution points such as goalkeeper claims, corners, or clearances. This will preserve tactical realism whilst still isolating the defensive mechanisms that genuinely end pressure phases.

Nonetheless, these shortcomings highlight rather than undermine the potential of recurrent approaches. The GRU model represents an early prototype for a broader methodological shift: treating football not as a collection of isolated events but as a series of structured, semi-deterministic chains. Such architectures could be repurposed beyond long throws to analyse goalkeeper restarts, centre-back initiation patterns, or pressing triggers, in other words, any situation in which teams execute rehearsed, temporally linked actions. Moreover these 'rehearsed' actions may even be instinctive match phases - for example how once a winger goes past a fullback they are encouraged to cross or shoot. In this sense, the model's limitations reflect the difficulty of pioneering rather than the weakness of the concept itself.

The dataset covers five seasons and approximately 4,000 qualifying long throws, of which only around six per cent produced a shot and two per cent resulted in a goal. This extreme rarity introduces class imbalance and rare-event bias, complicating model calibration and inflating statistical uncertainty. In addition, the 360° frames often centre on the thrower, leaving parts of the penalty area outside the visible field. This reduces usable positional data, obscures marking assignments, and limits the accurate quantification of true defensive density. Although this was mitigated using different techniques for all the models (mentioned in their corresponding sections), future datasets with a more comprehensive coverage would help to overcome these blind spots by capturing complete defensive structures and movement patterns. Similarly, adopting expected goals (xG) values rather than binary shot or goal labels would mitigate the noise arising from rare outcomes. We also considered incorporating invented variables such as xDuelWin for the first aerial contest which would've been calculated given the existing data, but opted against it to focus on the other parts of the framework, nevertheless, such improvements could strengthen the model's early-phase resolution.

Finally, there remain conceptual tensions between the two modelling approaches. Long throws differ fundamentally from corners: they enter with lower velocity and often produce sustained, second-phase pressure even after an initial clearance. This characteristic made it difficult to synthesise the XGBoost model, which captures instantaneous risk at release, with the GRU model, which tracks extended temporal evolution in risk dynamics. The static model truncates danger too early, while the sequential model extends it too far. Bridging this divide, that is, capturing both the immediate spatial configuration and the evolving chain of events, in our opinion remains an open methodological challenge, but one that future hybrid architectures are well placed to address.

In sum, while the current GRU remains an early, imperfect prototype and the dataset is sparse, the combined framework marks a decisive methodological step. It shows that even chaotic restarts like long throws can be decomposed into measurable geometry and reproducible temporal logic.

Thus, this contribution brings football analysis one step closer to deterministic understanding of set-piece danger, that is, proving that even the chaos of a long throw follows underlying, measurable rules.

9. Conclusion

This paper set out to quantify and explain the defensive mechanisms that determine whether a long throw-in evolves into danger or dissolves into control. By combining interpretable feature modelling (XGBoost), temporal sequence learning (GRU), and systematic counterfactual testing, this paper provides one of the first data-driven frameworks focused explicitly on the defence of long throws rather than their attack.

Across all analytical layers, a coherent tactical narrative emerges. Crucially, geometry sets the ceiling of risk: every metre deeper that the throw originates from yields roughly a 5% reduction in shot probability. Moreover, our analysis reveals that the central corridor around the penalty spot remains the critical zone to defend. Time then decides the outcome: the two-to-six-second window following first contact is critical in defining whether defensive stability is restored or collapses. Within this short period, compact spacing, directional clearances, and pre-scripted recovery chains exert greater influence on modelled concession risk than raw player height, goalkeeper aggression, or numerical density. This means that the data thus redefine effective long-throw defence not as an act of physical or numerical dominance, but of disciplined coordination in space and time.

Methodologically, the project advances football analytics along two complementary frontiers. The XGBoost model demonstrates that interpretable geometric baselines can generate actionable tactical rules from static events and 360 data, while the GRU prototype extends analysis into the temporal domain thus showing that defensive behaviour can be modelled as evolving sequences rather than isolated moments. Though preliminary and imperfect, this sequential framing represents a crucial step toward a simulation-ready understanding of defensive behaviour: one that can eventually test not only what happened, but what *might* have happened under alternative structures or decisions if the counterfactual framework we suggested in Section 6 is followed.

Practically, the findings translate into clear coaching guidance:

1. Do everything possible to avoid conceding throw-ins closer to the goal.
2. Anchor centrally with five to six defenders, prioritising spacing over numbers.
3. Maintain defensive compactness for at least six seconds post-contact.
4. Clear directionally toward the sidelines near the halfway line not vertically or laterally.
5. Adopt patient and conservative keeper positioning, favouring depth and reaction time.

Adhering to these principles would reduce shot probability considerably thus turning one of football's most chaotic restarts into a (somewhat) controllable defensive phase.

Finally, this paper positions long-throw defence as a prototype domain for a broader research movement. That being the quantitative decoding of structured yet stochastic phases of play. The same architecture could be redeployed to goal-kick build-ups, pressing triggers, or transition chains, gradually transforming instinctive on-field decision-making into explainable sequential dynamics. In that sense, the long throw is not an endpoint but a methodological test case which demonstrates that even football's scrappiest sequences can be rendered interpretable, optimisable, and ultimately coach-actionable through the fusion of data, sequence learning, and tactical reasoning.

References

[1] Stone, J. A, Smith, A., & Barry, A. (2021) 'The undervalued set piece: Analysis of soccer throw-ins during the English Premier League 2018-2019 season.' International Journal of Sports Science & Coaching
[2] Casal et al. (2023) 'Effects of tactical dimension and situational variables in throw-ins on the offensive performance in football'