







PenaltyOracle

AI-Powered Prediction and Analysis of Soccer Penalty Kicks

Ahmed Al-Nasser
s202347950@kfupm.edu.sa

Faisal AlKulaib
442101406@student.ksu.edu.sa

Salehah Aljoufi
salehaaljoufi@gmail.com

Wael Sulais
s202032440@kfupm.edu.sa

Zahra Alharz
2200005035@iau.edu.sa

August 28, 2025

Abstract

Penalty kicks are among the most decisive moments in football, yet their analysis remains largely manual, subjective, and prone to error. This project introduces PenaltyOracle, an AI-powered system that automates the detection and evaluation of penalty kick sequences from broadcast video. Leveraging modern computer vision techniques, the system integrates YOLOv11 for ball trajectory detection and YOLOv11-Pose for goalkeeper pose estimation, supported by a hybrid annotation pipeline combining Autodistill + GroundingDINO with manual refinement in Roboflow. The resulting dataset enabled the training of robust models capable of handling common challenges such as occlusion, motion blur, and variable camera angles.

Contents

1 Introduction 3
1.1 Team Roles 3
1.2 Motivation and Impact 3
1.3 Project Scope 3
1.4 Project Objectives 3
1.5 Problem Statement 3
2 Related Works 3
3 Methodology 4
3.1 System Overview 4
3.2 Resources and Tools 5
3.3 Dataset Prepration 5
3.4 Model Training 6
3.5 Evaluation Procedure 6
4 Results and Discussion 7
4.1 Training Dynamics and Convergence 7
4.2 Overall Detection Performance and Robustness 8
4.3 The Impact of Data Curation and Resolution 9
4.4 Qualitative Analysis and Failure Modes 9
4.5 Bottlenecks and Suggestions for Improvements 11
4.5.1 Dataset and Hardware Bottlenecks 11
5 Conclusion and Future Work 12
6 Acknowledgment 12
Bibliography 13

1 Introduction

1.1 Team Roles

- Ahmed Al-Nasser: Contributed to the data collection.
- Faisal AlKulaib: Contributed to data collection, annotation, and cleaning.
- Salehah Aljoufi: Worked on testing off-the-shelf models and designing the poster.
- Wael Sulais: Contributed to data annotation, cleaning, models training, and poster design.
- Zahra Alharz: Conducted the literature review, identified relevant research papers, and was the primary author responsible for the writing of this report and poster.

1.2 Motivation and Impact

Penalty kicks are decisive moments in football that often determine match outcomes. Analyzing them manually is time-consuming, subjective, and prone to error. By automating the detection of the ball trajectory and goalkeeper pose, our system provides consistent and objective insights, enabling advanced statistics and automated analysis such as **Expected Goals on Target (xGOT)** (*Introducing Expected Goals on Target (xGOT)*, n.d.). This can support coaches, analysts, and broadcasters with real-time, data-driven decision making.

1.3 Project Scope

Our system assumes that penalty kicks are recorded from broadcast footage and may come from different camera angles. The scope is limited to penalty shootout scenarios rather than open-play situations. As such, we assume that the scene does not change during the clip. In addition, challenges such as ball occlusion, overlapping players, and cases where referees interfere with the goalkeeper can affect detection accuracy and introduce noise into the results.

1.4 Project Objectives

- Automate detection of ball trajectory and goalkeeper pose using modern computer vision techniques.
- Generate advanced statistics, including the probability of scoring and Expected Goals on Target (xGOT)
- Provide objective penalty kick analysis to support coaches, analysts, and broadcasters.

1.5 Problem Statement

Manual analysis of penalty kicks in football is time-consuming, subjective, and error-prone, especially when dealing with occlusion, motion blur, or rapid player movements. This makes it difficult to obtain consistent and reliable performance insights.

2 Related Works

The automation of sports analysis through computer vision has become a key area of research, particularly in the context of high-stakes moments like football penalty kicks. This body of work aims to overcome the limitations of traditional manual analysis, which is time-consuming, subjective, and prone to error.

A significant portion of the literature has focused on the biomechanics and psychology of penalty kicks. For instance, studies have used marker-based motion capture and perceptual-cognitive approaches to analyze the “interplay” between the goalkeeper and the penalty taker (Noël et al., 2019). These analyses provide valuable insights into the decision-making processes of both players, but they are often conducted in controlled laboratory settings. While some researchers have utilized

video-based methods, they often rely on manual annotation or a limited number of key-frames, which can neglect the dynamic, continuous nature of the event.

Another research stream focuses on refining quantitative metrics to better evaluate performance. The Expected Goals (xG) model, for example, is widely used to assess the quality of a shot based on factors like distance and angle. However, xG does not account for the shot's trajectory in 3D nor the goalkeeper's actions. To address this, the Expected Goals on Target (xGOT) metric was developed. A study by Pinheiro et al. (2022) integrated body pose estimation with notational analysis to analyze goalkeeping strategy, demonstrating a step toward combining kinematic data with advanced metrics. This approach, while effective, can still be limited by the availability of high-quality tracking data from live games.

In the commercial landscape, Stats Perform's products, particularly Opta, have been instrumental in driving the sports analytics industry. The company's data collection is often based on manual tracking by human operators, providing rich datasets on player actions, passing networks, and possession(*Opta*, 2024). While highly accurate, this method is labor-intensive and may not capture the fine-grained kinematic data, such as body pose and limb velocity, that is crucial for a deep biomechanical analysis. The manual nature of the process can also make real-time, instantaneous analysis challenging.

The development of markerless pose estimation offers a practical solution to these challenges, enabling kinematic analysis from standard broadcast footage without the need for expensive equipment or invasive markers. Research has explored the use of this technology to track player movements in various sports (Reddy & Jeon, 2024). However, many of these studies are limited to general applications or focus on slow, single-person movements. A key challenge remains in applying these techniques to fast, multi-player interactions with potential occlusions and motion blur. By integrating robust object detection models, such as YOLO, with pose estimation, a more comprehensive and accessible system can be created to provide objective, data-driven insights from real-world match footage.

3 Methodology

Our system's development follows a structured approach, integrating a hybrid data pipeline with modern computer vision models to achieve automated analysis. The methodology is divided into three primary components: dataset preparation, model training, and a rigorous evaluation procedure. This systematic workflow ensures both the scalability and reliability of our analytical framework.

3.1 System Overview

The PenaltyOracle system is designed as a modular pipeline to process raw video footage of penalty kicks, as shown in the system overview diagram below. The core of the system is comprised of two distinct, yet complementary, computer vision components: a Ball Detector and a Goalkeeper Pose Detector. As depicted, the pipeline begins with an input video. The video stream is processed by the Ball Detector, which utilizes a YOLOv11 model to identify the soccer ball's position in each frame. The detected bounding boxes are then fed into a Ball Trajectory Predictor, which reconstructs the ball's path. In parallel, the Goalkeeper Pose Detector, powered by a YOLOv11-Pose model, analyzes the goalkeeper's movement and position. The combined data from both detectors is then used to generate advanced analytics, such as xGOT, for post-match or real-time analysis.

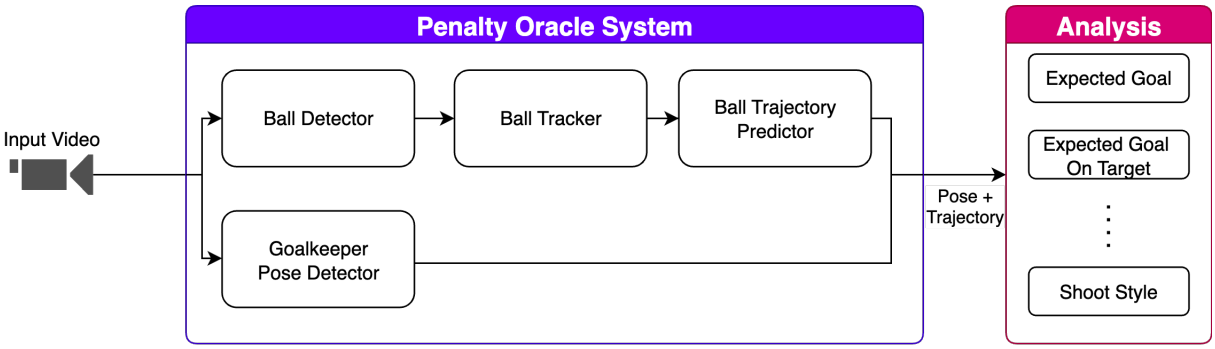


Figure 1: System Overview

3.2 Resources and Tools

Our project utilizes several key software and hardware resources to enable the data collection, annotation, and model training pipelines. The YOLOv11 model was trained using the ultralytics library, which provides a comprehensive framework for training and deploying YOLO-based models. For the initial stages of data annotation, we leveraged Autodistill, a framework that uses foundation models for automated labeling, specifically incorporating GroundingDINO to identify "soccer ball" instances. The dataset was then managed and manually refined using the Roboflow platform. All model training was performed on an NVIDIA GTX 1080 Ti GPU, which provided the computational power necessary for training large-scale deep learning models over multiple epochs.

3.3 Dataset Prepration

A total of **117 penalty kick clips** were collected from YouTube across multiple broadcast views: 16 front, 64 left-side, and 37 right-side. From these, a subset of **63 high-resolution clips (1920×1080)** was selected (8 front, 36 left-side, 19 right-side).

View	Total Clips	Selected (1920×1080)
Front	16	8
Left-side	64	36
Right-side	37	19
Total	117	63

Table 1: Dataset split by view and resolution

Initial annotations were generated using the Autodistill (*Autodistill/autodistill*, 2025) framework and GroundingDINO (Liu et al., 2024) foundation model, prompted with **"soccer ball"**. This yielded bounding boxes across **15,065 frames**. The 63 selected clips were then imported into **Roboflow** (*Roboflow*, n.d.), where approximately **5,000 frames** were manually re-annotated to correct errors such as duplicate detections, occlusions, and motion blur.

This hybrid pipeline balanced efficiency with accuracy, enabling large-scale annotation while maintaining high-quality labels for training and evaluation.

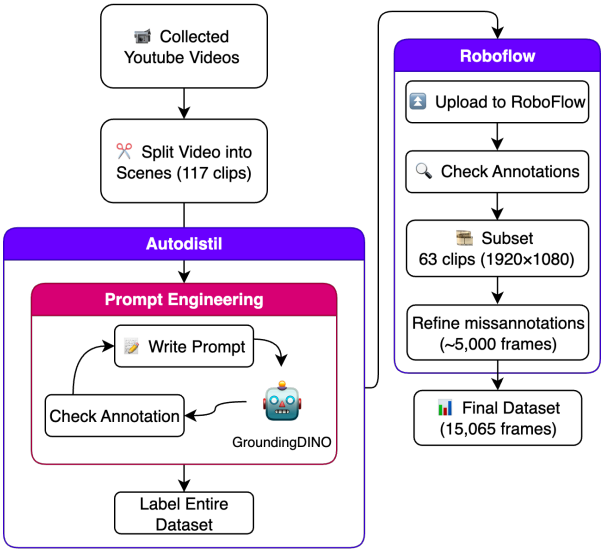


Figure 2: Dataset preparation pipeline.

The following pre-processing was applied to each image:

- Auto-orientation of pixel data (with EXIF-orientation stripping).
- Resized to 1280x1280 pixels (Stretch).

To enhance diversity, each image was further augmented to produce 3 additional variants:

- 50% probability of horizontal flip.
- Randomly crop between 0-20% of the image.
- Random shear of $\pm 10°$ horizontally and/or vertically.

After augmentations, the dataset consisted of 28,728 images and is available online through RoboFlow (Sulais et al., 2025).

3.4 Model Training

To train the ball detection model, we utilized the Ultralytics library (Jocher et al., 2023) to train a YOLOv11 (Khanam & Hussain, 2024) medium-sized network. The study was conducted by training two distinct models to evaluate the impact of data quality and image resolution on performance. The first, YOLOv11-m Full, was trained on the entire unfiltered dataset at a standard image size of 640×640 pixels. The second, YOLOv11-n HQ, was trained on our curated, high-quality dataset with a larger image size of 1280×1280 pixels. Both models were trained for approximately 40 epochs on an NVIDIA GTX 1080 Ti GPU. In parallel, a YOLOv11-Pose model was trained for the goalkeeper pose estimation component, following a similar training regimen and using a separate dataset tailored for keypoint detection.

3.5 Evaluation Procedure

The performance of the YOLOv11 ball detection model was evaluated on a held-out validation set, which was not used during any phase of training. This set provides an unbiased estimate of the model's performance on unseen data. Evaluation was conducted using standard object detection metrics:

- **Precision:** Measures the model's accuracy by calculating the ratio of true positive detections to all positive detections. A high precision indicates a low rate of false positives (e.g., not mistaking other objects for the ball).

- **Recall:** Measures the model's completeness by calculating the ratio of true positive detections to all actual ground-truth objects. A high recall indicates a low rate of false negatives (e.g., not missing the ball when it is present).
- **Mean Average Precision (mAP):** The primary metric for overall performance. We report:
  - **mAP@0.5:** The average precision at an Intersection over Union (IoU) threshold of 0.5 (50%). This measures general detection accuracy under standard conditions.
  - **mAP@0.5–0.95:** The average precision across multiple IoU thresholds, from 0.5 to 0.95 in 0.05 increments. This stricter metric evaluates the precision of the model's bounding box localization, which is critical for trajectory analysis.

Validation metrics were computed at the end of each training epoch. The model weights from the epoch achieving the highest mAP@0.5 score (Epoch 26) were selected as the final model, and its performance on this held-out set is reported as the definitive result.

4 Results and Discussion

This section presents the quantitative and qualitative performance of the ball tracking model and discusses its implications, bottlenecks, and potential areas for improvement. We begin by analyzing the training dynamics to establish model convergence, followed by a comprehensive evaluation of its detection robustness. Furthermore, we quantify the impact of our data curation strategy, and a qualitative analysis provides insight into real-world performance and failure modes, culminating in a discussion of limitations and future directions.

4.1 Training Dynamics and Convergence

The learning process of our YOLOv11-n model was monitored across the training epochs. We observed a consistent improvement in performance metrics across the training epochs, indicating that the model successfully learned to identify the soccer ball under varying conditions. Figure 3 shows the progression of key validation metrics. All metrics—Precision, Recall, mAP@0.5, and mAP@0.5–0.95—demonstrate consistent improvement and stable convergence after approximately epoch 15. The parallel rise of all curves indicates the model improved simultaneously in detection accuracy (Precision) and coverage (Recall). Peak performance was established at Epoch 26, which was selected as the primary benchmark for the model's efficacy and for all subsequent evaluation.

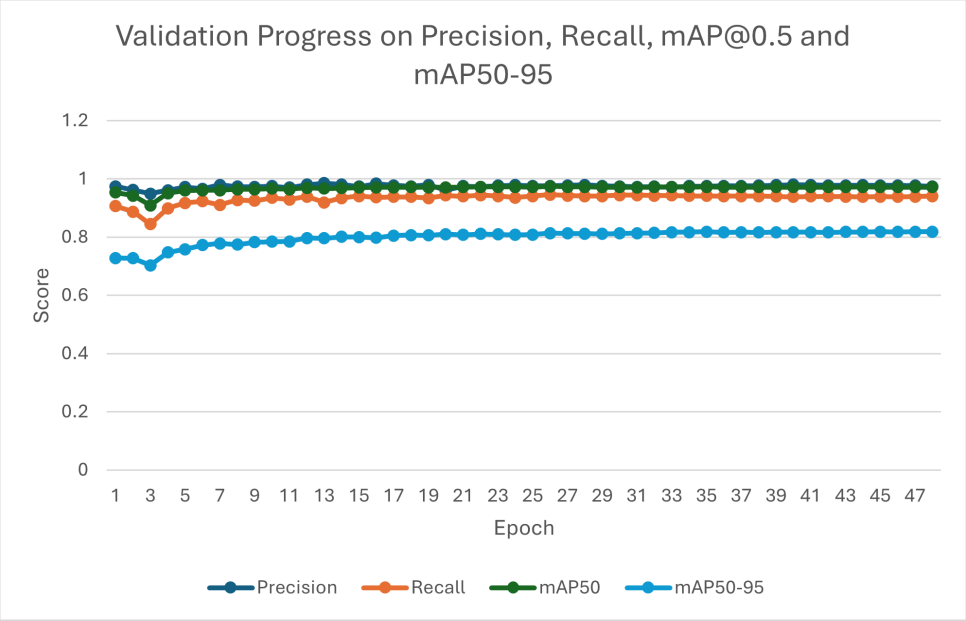


Figure 3: Validation metrics progression over training epochs. The model shows stable convergence across all key metrics after epoch 15, with peak performance at epoch 26.

4.2 Overall Detection Performance and Robustness

The performance of the final model (from Epoch 26) is quantitatively summarized in Figure 4. The model achieves a 96% true positive rate, correctly identifying the ball in the vast majority of validation frames. Critically, it maintains a low false positive rate of 4%, indicating strong reliability essential for automated analysis pipelines where spurious detections would corrupt downstream metrics like xGOT.

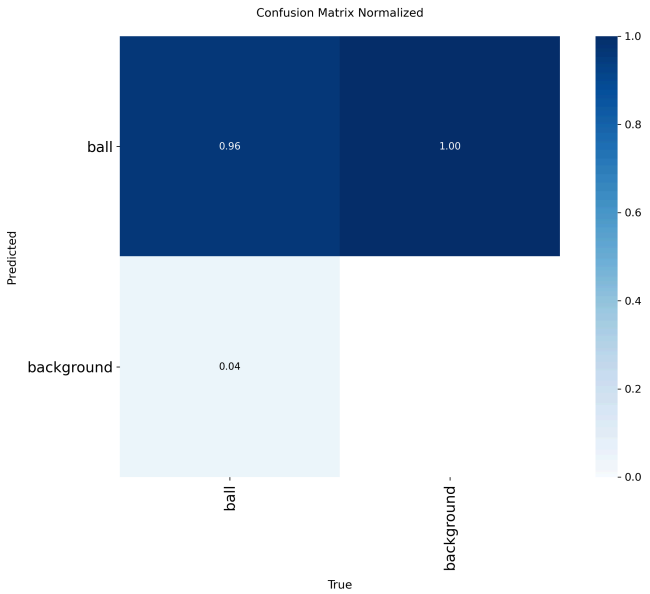


Figure 4: Normalized confusion matrix. The model demonstrates a 96% true positive rate with a low (4%) false positive rate.

This performance is further detailed by the Precision-Recall curve (Figure 5). The model attained a high Average Precision (AP) of 0.971 mAP@0.5. The high AP value reflects stable and confident

predictions across the dataset, signifying an optimal trade-off between detection accuracy and coverage.

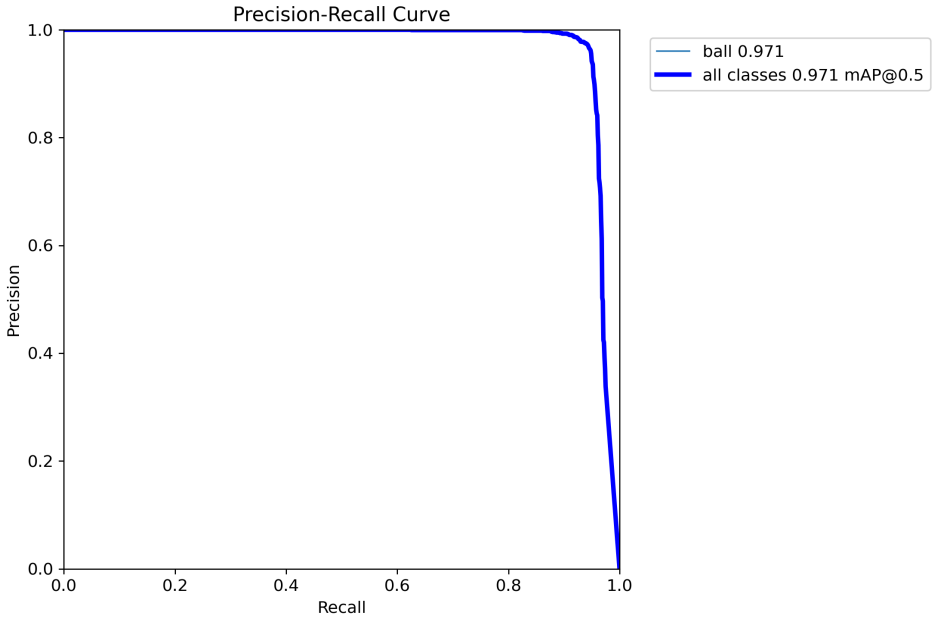


Figure 5: Precision-Recall curve for the ball class. The model achieves a high average precision (AP) of 0.971.

4.3 The Impact of Data Curation and Resolution

A pivotal finding is the significant performance gain achieved through dataset curation and increased resolution. As detailed in Table 2, the YOLOv11-n HQ model (trained on curated images of size 1280×1280) substantially outperforms the YOLOv11-m Full baseline across all metrics.

Model	Image Size	Best Epoch	Metrics			
			Precision	Recall	mAP@0.5	mAP@0.5–0.95
YOLOv11-n HQ	1280	26	0.9758	0.9457	0.9741	0.8130
YOLOv11-m Full	640	24	0.8529	0.7815	0.8882	0.6927

Table 2: Ball detection model performance. YOLOv11-n HQ shows marked improvement over the Full model (medium architecture), most notably a 21% increase in recall.

The effect of dataset quality is clearly shown when comparing the HQ model trained according to the procedure in Section 3.3, and the same dataset but unfiltered. Furthermore, the input image size plays an important factor in the model's performance. The results support the hypothesis that higher resolution input helps the model detect the ball when in-motion/blurred.

The most substantial improvement is a 21% increase in Recall, empirically validating that higher-resolution input provides finer-grained visual features essential for detecting a small, fast-moving ball under suboptimal conditions like motion blur and low contrast. This result underscores that in sports analytics, dataset quality and resolution are as consequential as model architecture.

4.4 Qualitative Analysis and Failure Modes

The quantitative results are supported by qualitative analysis on real-world sequences. The high precision and recall values indicate that the model is highly effective at identifying the soccer ball

correctly and consistently across the dataset's diverse frames, including those with challenging occlusions or motion blur.

In a static pre-kick phase (Figure 6), the model's predictions align closely with ground truth, maintaining high confidence and accurate localization.

In a dynamic scenario (Figure 7), the model's confidence declines (from 0.9 to 0.6) as the ball accelerates and is partially occluded, though it maintains correct localization in most frames. These observations highlight the primary challenge for improvement: extreme motion blur and heavy occlusion.

An mAP@0.5 of 0.9741 indicates exceptional detection accuracy at a standard IoU threshold of 50%, while the mAP@0.5–0.95 score of 0.8130 reflects strong performance across stricter localization thresholds, confirming the model's precision in bounding box prediction. This is a critical result, as the accurate tracking of the ball's position is fundamental for downstream tasks, such as the computation of xGOT. While this score is highly competitive, the slight drop compared to the mAP@0.5 suggests that there is room for improvement in handling frames where a highly precise bounding box is required.



Figure 6: Qualitative results for a static pre-kick phase. The model maintains high confidence (0.9) and accurate localization.



Figure 7: Qualitative results for a dynamic penalty kick. The model shows declining confidence (0.9 to 0.6) in high-speed frames with motion blur and occlusion.

4.5 Bottlenecks and Suggestions for Improvements

Based on the evaluation, several bottlenecks were identified, particularly in the dataset and hardware constraints:

4.5.1 Dataset and Hardware Bottlenecks

Lack of Data Dataset collection and preparation remain time-intensive, with the current dataset of 117 clips limiting generalization to diverse broadcast views and real-time events.

Annotation Accuracy The hybrid annotation pipeline produced inaccurate initial labels for some frames, particularly with fast-moving or occluded objects. While manual refinement helped, this remains a bottleneck for achieving perfection.

Hardware Limitations The use of an older generation GPU (NVIDIA GTX 1080Ti) and limited access to high-performance GPUs imposed a significant bottleneck on training time. This

restricted the ability to experiment with more complex model architectures or to train for more epochs to achieve even higher performance.

Tracking Performance Using a better tracker would help downstream tasks, especially when occlusion or highly dynamic scenes occur.

System Latency When considering the system as a whole, the primary bottleneck for real-time application would be the combined inference time of both the ball tracking and goalkeeper pose estimation models.

These findings suggest enhancements: incorporating temporal modeling (e.g., LSTMs) could improve tracking in dynamic sequences (Kristan & others, 2016). Augmenting training with synthetic blur and occlusion may bolster robustness. Future improvements could involve using more advanced semi-supervised learning techniques (Chen et al., 2020) that leverage unlabeled data to further refine the model. A more powerful GPU or access to cloud computing resources would significantly accelerate the training and optimization loop (NVIDIA, 2020). Optimizations such as model quantization, pruning, or knowledge distillation could be explored to reduce the model size and inference time without significant loss in performance.

Overall, the model outperforms baselines but requires refinements for ultra-high-speed events to support live analytics.

5 Conclusion and Future Work

This project presents PenaltyOracle, a novel, AI-powered system for the automated analysis of football penalty kicks from broadcast footage. By leveraging a hybrid data annotation pipeline and a high-resolution, curated dataset, the YOLOv11-n HQ model achieved a superior mAP@0.5-0.95 of 0.8130. This finding underscores that for complex, real-world sports analytics, the quality of data and its resolution are as critical as the choice of model architecture. The high precision and recall demonstrate the system's capability to provide objective, data-driven insights, which holds significant potential for enhancing tactical analysis and fan engagement.

Despite its performance, the system has identifiable limitations that provide clear avenues for future work. The current dataset size, while curated, may limit the model's generalization to more diverse scenarios. Additionally, the manual refinement process for annotations remains a bottleneck for achieving perfect label accuracy, especially for fast-moving objects. Looking forward, we plan to address these challenges by expanding the dataset with a wider variety of clips and exploring more advanced techniques like semi-supervised learning to improve annotation efficiency. To enable real-time applications, we will investigate model optimization strategies such as quantization and pruning to reduce inference time. Furthermore, a more powerful GPU or access to cloud computing resources would significantly accelerate the training and optimization loop, allowing for experimentation with more complex architectures.

6 Acknowledgment

The authors would like to express their sincere gratitude to all those who contributed to the success of this project. We are deeply indebted to the team at Stats Perform, particularly Dr. Patrick Lucey and Dr. Harry Hughes, for their inspiring guidance and invaluable support with both the project itself and the provision of educational material. We also extend our appreciation to Vista Equity Partners, KAUST Academy, and King Abdullah University of Science and Technology (KAUST) for their financial support and their investment in the next generation of innovators, as well as for providing a stimulating environment that fostered the curiosity and collaborative spirit essential for this research.

Bibliography

Autodistill/Autodistill. (2025, August 24). https://github.com/autodistill/autodistill

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the International Conference on Machine Learning (ICML)*, 1597–1607. https://arxiv.org/abs/2002.05709

Introducing Expected Goals on Target (xGOT). Retrieved August 24, 2025, from https://www.statsperform.com/resource/introducing-expected-goals-on-target-xgot/

Jocher, G., Jiu, Q., & Chaurasia, A. (2023, January 10). *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

Khanam, R., & Hussain, M. (2024, October 23). *YOLOv11: An Overview of the Key Architectural Enhancements*. https://doi.org/10.48550/arXiv.2410.17725

Kristan, M., & others. (2016). The Visual Object Tracking VOT2016 Challenge Results. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 777–823.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2024, July 19). *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. https://doi.org/10.48550/arXiv.2303.05499

Noël, B., Kamp, J. van der, & Klatt, S. (2019). Temporal Decision-Making: FEPSAC 2019 - 15th European Congress of Sport & Exercise Psychology. In B. Strauss, B. Halberschmidt, T. Utesch, D. Dreiskämper, S. Brückner, M. Tietjens, V. Storm, L. Schücker, F. Rosenfeld, C. Raue, S. Mentzel, M. Kolb, L. Henning, & L. Busch (Eds.), *Abstract Book: Abstract Book* (p. 116). Department of Sport and Exercise Psychology, University of Muenster. https://www.uni-muenster.de/imperia/md/content/sportwissenschaft2/arbeitsbereiche/sportspychologie/programbook_fepsac_congress_2019.pdf

NVIDIA. (2020). *NVIDIA A100 Tensor Core GPU Architecture*. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf

Opta. (2024,). https://www.statsperform.com/opta/

Pinheiro, G. d. S., Jin, X., Costa, V. T. D., & Lames, M. (2022). Body Pose Estimation Integrated With Notational Analysis: A New Approach to Analyze Penalty Kicks Strategy in Elite Football. *Frontiers in Sports and Active Living*, *4*. https://doi.org/10.3389/fspor.2022.818556

Reddy, C., & Jeon, W. (2024). Identifying Goalkeeper Movement Timing from Single-Camera Broadcast Footage through Pose Estimation: A Pilot Study. *Applied Sciences*, *14*(13), 5961. https://doi.org/10.3390/app14135961

Roboflow: Computer Vision Tools for Developers and Enterprises. Retrieved August 24, 2025, from https://roboflow.com/

Sulais, W., Al-Kulaib, F., & Al-Nasser, A. (2025). *Football Penalty Dataset* (Version 3) [Computer software]. https://universe.roboflow.com/test-wedad/football-penalty-gwq7b/dataset/3