Review-Informed Pilot Dataset and Baseline Modeling for NBA Game Winner Prediction

Abdulla Ali, Masih Faryadi, Yoseyas Surafel, Zaid Abu-Abbas, and Xiaodong Qu
Department of Computer Science, The George Washington University
Washington, DC, USA
{dulee,masih22,ysurafel4,zabuabbas23,x.qu}@gwu.edu

Abstract

Predicting the winner of an NBA game is a useful pilot problem for studying how sports data, feature construction, and machine learning interact in a realistic data mining setting. This paper presents a review-informed student pilot study focused on pre-game NBA winner prediction. Building on an earlier review-oriented project, we organize the work around three steps: summarizing prior sports prediction ideas that motivate the empirical setup, cleaning and structuring NBA API-derived game data into pilot datasets, and comparing baseline machine learning models for pre-game prediction. The dataset work centers on raw NBA game logs and engineered matchup-level features such as team statistical differentials, recent performance indicators, and pre-game strength estimates. The experiments are treated as preliminary rather than as a completed benchmark. In the reported pilot results, a Transformer model has the strongest predictive metrics, while SVM remains a competitive and efficient baseline. The goal is to document a concrete direction for student-scale sports analytics research while being clear about remaining limitations, including missing code documentation, uncertain split details, and the need to verify feature timing before making stronger claims.

Keywords

sports analytics, NBA prediction, machine learning, explainable AI, student research

ACM Reference Format:
Abdulla Ali, Masih Faryadi, Yoseyas Surafel, Zaid Abu-Abbas, and Xiaodong Qu. 2026. Review-Informed Pilot Dataset and Baseline Modeling for NBA Game Winner Prediction. In *Proceedings of KDD 2026 Undergraduate Consortium (KDD UC '26).* ACM, New York, NY, USA, 6 pages.

1 Introduction

Sports prediction is a natural data science problem because the outcome is easy to define but difficult to explain. In professional basketball, a game winner depends on long-term team quality, current form, matchup context, roster changes, injuries, coaching decisions, and many other factors that shift over the course of a season. These properties make NBA game prediction a useful setting for studying how students can move from a literature review to an empirical data mining workflow.

This paper focuses on pre-game NBA game winner prediction. Rather than claiming to build a finished prediction system or a benchmark dataset, we present a student-scale pilot study for the KDD Undergraduate Consortium. The project combines three parts: a review-informed framing of sports prediction, construction of cleaned NBA pilot datasets, and baseline pre-game prediction experiments using engineered game-level features. Building on an earlier review-oriented project, we move toward an empirical pilot study that can be checked, improved, and extended in later work. This review-to-pilot path follows prior lab-based undergraduate research efforts that use literature review, dataset organization, and structured evaluation to help students move from coursework toward reproducible data science research [23, 27, 31, 37, 38]. This study is part of the broader StatLine student project, which aims to combine NBA game data, predictive modeling, and interpretable explanations in a single sports analytics prototype.[1]

Pre-game prediction is intentionally narrower than live win-probability modeling. It allows the study to focus on information that should be available before tipoff, such as team statistics, recent performance, and pre-game team strength estimates. This scope is important because using post-game or future-game information would make the prediction task easier but would not reflect a realistic forecasting setting. In this pilot study, in-game prediction is not treated as a main experiment because the evidence gathered for this submission does not yet include clearly separate in-game datasets and result files.

The paper is organized around three contributions:

(1) A review-informed framing of NBA and sports game prediction that motivates the use of pre-game statistical features and baseline model comparison.
(2) A cleaned and organized pilot dataset built from NBA API game data, including raw season game logs and engineered matchup-level features for NBA game winner prediction.
(3) Baseline pre-game prediction experiments comparing multiple machine learning models in the current pilot setup, with model claims stated as preliminary and tied to the available data files.

2 Background and Related Work

Prior work on basketball outcome prediction often frames the problem as a pre-game supervised learning task. Earlier student and course-project studies used NBA API-derived data and classical models such as logistic regression and SVMs to predict game outcomes from season statistics and matchup features [10, 26, 29]. These studies are useful for the present paper because they show how a prediction task can be defined at the game level and evaluated with relatively compact tabular features.

Basketball prediction research also includes neural-network models, data mining classifiers, statistical forecasting models, and hybrid data-mining schemes for NBA outcomes [4, 15, 18, 20]. Across team

KDD UC '26, Jeju, Korea
2026.

[1] A preliminary student prototype is available at: http://52.70.223.60/

sports, rating-based approaches such as Elo further show how pre-game strength estimates can be converted into predictive covariates [9].

More recent NBA prediction work emphasizes the importance of feature engineering and model comparison. Studies using feature analysis and machine learning identify box-score and efficiency indicators, recent form, and team-level statistical summaries as important inputs for outcome prediction [8, 33, 36]. Broader reviews of sports prediction and basketball prediction also caution that reported performance depends heavily on data quality, league context, feature design, and evaluation protocol [2, 13]. This is why we frame the work as a pilot study rather than a completed benchmark. Related undergraduate-oriented work in another human-data domain has likewise used structured literature review to organize machine learning methods, datasets, and research opportunities before focused empirical investigation [22].

Recent empirical work also shows that stronger models can take different forms depending on how the basketball problem is represented. Benchmarking and stacked ensemble studies compare multiple learning algorithms across NBA outcome settings [1, 7], while graph-based approaches model team relationships and league structure directly [39, 40]. These directions motivate comparing a range of model families in this pilot study while still treating the results as dataset- and split-dependent. Recent work has also explored hybrid neural architectures for NBA playoff outcome prediction [11], extending basketball modeling beyond conventional tabular classifiers. Related studies in other human-data domains have likewise explored Transformer-based and hybrid convolution–Transformer architectures for structured prediction tasks [14, 41].

Basketball outcome prediction has also been studied as a sequential or state-based problem, especially for win probability estimation during games [12, 19, 30, 32, 35]. Those studies help explain why in-game modeling is an important future direction, but the present paper intentionally narrows the main empirical task to pre-game prediction because separate in-game datasets and results are not yet documented for this submission.

The project is also informed by work connecting prediction with interpretability. Feature-attribution and local-explanation methods such as SHAP and LIME are widely used to connect model outputs to input features, but their explanations should be interpreted as model behavior rather than causal evidence [6, 17, 28]. Human-centered explanation research also emphasizes that explanations should be understandable to the people using the prediction, not only technically available [21]. Ouyang et al. [24] combine XGBoost and SHAP for NBA outcome prediction and show how feature contributions can be used to explain model behavior. Sports visualization and explainable analytics work also emphasizes that prediction tools are more useful when outputs can be connected to domain concepts that analysts and users recognize [5, 25, 34]. Here, feature contribution evidence is treated as exploratory because the current experimental notes do not fully document the source model or SHAP procedure. Together, these studies motivate a review-informed empirical setup: construct inspectable pre-game features, compare several baseline models, and avoid broad claims until the data pipeline and evaluation strategy are fully documented.

3 NBA Data Collection and Pilot Dataset Construction

The empirical part of this project begins with NBA game data collected through the NBA API and saved as season-level game logs. The dataset audit identified raw game-log CSV files organized by season, with fields for team identity, game date, matchup, win/loss outcome, shooting statistics, rebounds, assists, turnovers, points, and scoring margin. These files provide the raw material for constructing pre-game prediction examples.

3.1 Data Processing Workflow

The data-processing workflow begins with season-level NBA game logs derived from NBA API data. These files are organized as team-level records, so the first processing step is to clean and align team and game identifiers across season files. This alignment is important because a prediction example should represent a single game, not two unrelated team rows.

After alignment, the workflow converts team-level logs into game-level matchup rows. Each matchup row is organized around the home team and away team, with a target label indicating whether the home team won. The engineered features are then constructed as home-minus-away differentials, so the model input captures relative matchup advantages in recent performance, scoring and efficiency summaries, and pre-game strength estimates.

The same workflow supports both the multi-season historical engineered dataset and the same-season engineered dataset summarized in Table 1. These model-ready rows are then used for the pilot model comparison. The study does not yet include released scripts for every processing step, and the timing of rolling, blended, and rating-based features still needs confirmation. A future version should include the full feature-generation code and clearer documentation of how each pre-game feature is computed.

The engineered feature files use home-minus-away differences and include blended shooting, scoring, rebounding, assist, turnover, plus-minus, and net-margin differences, along with a recent win percentage difference and a pre-game Elo difference.

Two engineered pre-game feature files appear to support different pilot settings. One file is larger and includes season-tracking columns, while the other is smaller and appears to correspond to a more recent or updated pre-game dataset. This distinction is potentially important because the student notes and Lab 9 materials discuss comparing recent or same-season data against a broader historical setting. At this stage, the paper should describe that comparison cautiously until the exact definitions of the datasets are confirmed.

Table 1 summarizes the raw logs and the two engineered feature datasets used in the pilot study. Early raw files may include playoff rows, and the 2025–2026 data should be treated as a regular-season window through 2026-04-12. The exact definitions of the multi-season historical and same-season settings still require confirmation.

This dataset is best understood as a pilot dataset rather than a complete benchmark. The raw season files do not all expose the exact same set of columns, and this study does not yet include the code used to download data, build features, or reproduce the train/test split. These limitations do not prevent the dataset from

NBA API game logs → cleaning and schema alignment → game-level matchup rows → pre-game feature engineering → model comparison

Figure 1: Overview of the pilot data-processing workflow from NBA API-derived game logs to model-ready pre-game prediction examples.

Table 1: Pilot NBA data files available for the pre-game prediction study.

Dataset component	Rows	Cols.	Scope	Role
Raw game logs	38,916	32–34	2010–11 to 2025–26	Source season logs
Multi-season engineered pre-game	17,889	22	2011–12 to 2025–26	Historical feature set
Same-season engineered pre-game	1,209	20	2025–26 through 2026-04-12	Recent-season feature set

supporting a student pilot study, but they should be made explicit so that the paper does not overstate the maturity or reproducibility of the dataset.

4 Pre-Game Prediction Task and Experimental Setup

The prediction task in this pilot study is to estimate whether the home team will win an NBA game before the game begins. Each example corresponds to one scheduled matchup, and the label is the game winner represented as a home-team win outcome in the engineered data files. This framing keeps the experiment focused on pre-game forecasting rather than live win-probability modeling.

The model inputs are engineered pre-game features derived from NBA API game logs. As summarized in Table 1, the project includes raw season game logs, a multi-season historical engineered pre-game dataset, and a same-season engineered pre-game dataset. The raw logs provide the source material, while the engineered files organize each matchup into a single row for model comparison.

The feature representation is described at the group level because the exact feature-generation code is not yet part of the reproducibility materials. The engineered inputs capture team and matchup context, home/away context, recent performance indicators, scoring and efficiency summaries, and win/loss or rating-based summaries. These features are represented primarily as relative home-versus-away measures so that the model receives information about matchup differences rather than only isolated team statistics. The home-team win label is used for supervision and evaluation and is excluded from the model inputs. The engineered feature set includes 11 home-minus-away difference features, covering shooting, scoring, rebounding, assists, turnovers, margin, recent form, and Elo-style strength.

4.1 Feature Timing and Leakage Considerations

Because the task is pre-game prediction, each feature should be computed only from games that occurred before the target game. Rolling averages, recent win percentage, blended statistics, and Elo-style ratings should therefore be lagged with respect to the prediction game. Same-game final score, same-game box-score statistics, and the home-win label should not be included as model inputs. The current experimental notes do not yet fully document these timing checks, so the results should be treated as pilot evidence rather than a fully leakage-verified evaluation.

The pilot comparison uses the model list visible in the result file: SVM, CNN 1D, KNN, Random Forest, XGBoost, and Transformer. These models provide a mix of traditional machine learning baselines, ensemble methods, and neural architectures for the same pre-game prediction task. This section only defines the comparison setup; model performance is reported separately in the Results section.

The Lab 9 project materials describe the evaluation as an 80/20 split; however, the exact split dates and preprocessing steps still require confirmation from the final code pipeline. The result file reports accuracy, precision, recall, F1, and runtime. For this reason, we do not claim a fully time-aware split or leakage-free preprocessing in this version.

5 Pilot Results

This section reports the pre-game pilot results used for the paper. The results should be read as preliminary evidence from a student research workflow, not as a general ranking of NBA prediction methods. The present study does not yet include the code or run logs needed to fully reproduce the experiments.

5.1 Model Comparison

Table 2 summarizes the model comparison reported in the pilot result file. In this table, the Transformer model has the highest Accuracy, Precision, Recall, and F1 values. This result is useful as a pilot finding, but it should not be interpreted as evidence that Transformer models are generally best for NBA prediction. The result is tied to these feature files, split, and implementation choices.

SVM remains a competitive baseline in the same table. Its Precision is close to the Transformer value, and its reported runtime is substantially smaller. For a student-scale pilot study, this makes SVM useful as a simpler baseline even though it is not the top model on the reported predictive metrics.

5.2 Same-Season and Multi-Season Results

The experimental materials also contain a comparison between a larger multi-season historical setting and a same-season or recent-data setting. In that comparison, the same-season/recent setting reports higher Accuracy and F1 than the larger historical setting for SVM, Random Forest, XGBoost, and Transformer. Table 3 summarizes the available comparison for the models whose values are documented in the experimental record.

Table 2: Pre-game pilot model comparison from the reported result file.

Model	Acc.	Prec.	Rec.	F1	Runtime
SVM	73.3%	82.6%	70.9%	76.3%	0.15
CNN 1D	73.7%	80.8%	74.5%	77.5%	19.2
KNN	70.3%	79%	69.5%	74%	0.0031
Random Forest	72.3%	80.3%	70.8%	75.3%	0.43
XGBoost	68.6%	73.3%	74.3%	73.8%	0.18
Transformer	78.1%	82.7%	79.9%	81.3%	12

Table 3: Same-season/recent versus multi-season historical comparison from the pilot materials.

Model	Hist. Acc.	Hist. F1	Recent Acc.	Recent F1
SVM	0.665	0.711	0.733	0.763
Random Forest	0.665	0.690	0.723	0.753
XGBoost	0.651	0.709	0.686	0.738
Transformer	0.654	0.723	0.781	0.813

For example, the Transformer row increases from 0.654 Accuracy and 0.723 F1 in the historical setting to 0.781 Accuracy and 0.813 F1 in the recent setting. SVM similarly increases from 0.665 Accuracy and 0.711 F1 to 0.733 Accuracy and 0.763 F1. The exact definitions of the historical and recent settings still require confirmation, so this table is treated as pilot evidence.

This comparison suggests that recent season context may matter in this pilot setup. The interpretation should remain cautious because the exact definitions of the two settings, the split procedure, and the feature-generation steps need to be documented before making a stronger claim.

One possible explanation is that NBA team data can be non-stationary across seasons, reflecting the broader challenge of learning under concept drift [16]. Rosters, coaching decisions, injuries, playing style, and team strength can shift enough that older seasons may not represent the current season well. A larger historical dataset provides more examples, but it may also introduce distribution mismatch if relationships learned from older games do not transfer cleanly to newer matchups. The pilot comparison suggests a tradeoff between sample size and season relevance, while still requiring confirmation of the dataset definitions and split strategy.

5.3 Feature Contribution Analysis

The available feature-contribution file provides a small exploratory view of which variables contributed positively or negatively in an explanation analysis. The largest positive entries include Elo rating, field-goal percentage differential, recent momentum, net margin, and defensive rebounds. The negative entries include turnovers, personal fouls, free-throw attempts, and three-point attempts.

This analysis is useful for connecting the pilot models back to basketball concepts, but it should be treated as descriptive rather than causal. Following the explanation literature [6, 17], the feature directions are best read as model-level explanations of the current pipeline, not as evidence that changing one basketball statistic would cause a different game outcome. The study does not identify the exact source model, dataset split, or SHAP computation procedure for the feature-contribution file.

Table 4: Exploratory feature-contribution summary from the pilot materials.

Feature or group	Direction	Basketball interpretation
Elo rating	Positive contributor	Team-strength signal
Field-goal percentage differential	Positive contributor	Shooting efficiency
Recent momentum	Positive contributor	Recent form
Net margin	Positive contributor	Scoring-margin strength
Defensive rebounds	Positive contributor	Possession control
Turnovers	Negative contributor	Lost possessions
Personal fouls	Negative contributor	Foul discipline
Free-throw attempts	Negative contributor	Foul-pressure context
Three-point attempts	Negative contributor	Shot-profile variance

6 Discussion and Limitations

The pilot results suggest that pre-game NBA winner prediction is a useful student-scale data mining problem when the task is framed carefully. The model comparison table shows that several model families can be applied to the engineered matchup features, and the strongest predictive values in the reported table come from the Transformer model. This may indicate that a more expressive architecture can capture useful interactions among team strength, recent form, and efficiency summaries in this dataset. However, this interpretation should remain narrow. The reported results support a preliminary finding, not a broad conclusion that Transformer models are best for NBA prediction.

The SVM result remains important even though it is not the highest-performing model on the reported predictive metrics. SVM is competitive in the reported table and has a much smaller reported runtime than the Transformer model. For an undergraduate pilot project, a simpler and efficient baseline is valuable because it is easier to explain, easier to audit, and less dependent on model complexity. This makes SVM a useful comparison point for future versions of the study.

The same-season versus multi-season pattern is also promising but should be handled cautiously. The experimental record suggests that the same-season or recent-data setting performs better than the broader historical setting for the reported models. A plausible explanation is that NBA teams change substantially across seasons because of roster changes, coaching, injuries, and playing style. This study does not prove that explanation, and the exact dataset definitions still need confirmation, so the result is best described as preliminary evidence that recent season context may matter.

The feature contribution analysis helps connect the model outputs back to basketball concepts such as team strength, shooting efficiency, recent momentum, turnovers, and fouls. This is useful for interpretation, especially in a student research setting where the goal is not only to report metrics but also to understand why a prediction might be made. At the same time, the feature contribution file is exploratory. The paper does not fully document the source model, dataset split, or SHAP procedure used to produce those values.

Several limitations are central to this pilot study. First, the dataset should be described as a pilot dataset, not as a completed benchmark. The raw season files and engineered feature files are useful for a KDD UC pilot study, but they are not yet packaged with full data-generation documentation. Second, possible feature leakage is the most important technical risk. Although the Methods section

describes what a leakage-safe pre-game setup should require, the current experimental notes do not yet verify the timing of rolling, Elo, blended, and recent-performance features. If any of these values include same-game or future-game information, the reported performance would be overly optimistic.

Third, the evaluation procedure needs clearer documentation. The pilot study uses an 80/20 train-test split, but the current notes do not confirm whether this split was random, chronological, season-based, or otherwise constructed. A future time-aware evaluation, training on earlier games and testing on later games, would better match the forecasting setting [3]. Relatedly, the reproducibility materials do not currently include code for data collection, feature construction, training, preprocessing, or evaluation. This limits the ability to verify the results and should be addressed before making stronger claims.

Finally, the raw season files show some schema differences across years, and the 2025–2026 data should be treated as a regular-season window through 2026-04-12 rather than as postseason data. These issues do not invalidate the pilot study, but they reinforce why the findings are presented as early empirical evidence rather than as a finalized NBA prediction benchmark.

7 Conclusion

This paper presented a review-informed pilot study of pre-game NBA game winner prediction. The work connects prior sports prediction ideas to a student dataset-construction effort, organizes NBA API-derived game logs into pilot feature datasets, and compares several baseline models on a pre-game winner prediction task. In the reported table, the Transformer model has the strongest predictive metrics, while SVM remains a competitive and efficient baseline.

The main value of this work is not a final claim about which model is best for NBA prediction, but a clearer empirical starting point for student sports analytics research. Future work should document the full code pipeline, confirm that all engineered features are leakage-free, use a time-aware evaluation design, and extend the dataset with clearer season definitions and reproducible preprocessing steps. Beyond the pre-game study reported here, StatLine is also exploring live win-probability updates, plain-language explanations of feature contributions, and interactive question-answering grounded in prediction context.

References

[1] R. Alves and H. Barbosa. 2025. Benchmarking machine learning models for NBA and WNBA game outcome prediction. *Journal of Sports Analytics* (2025).

[2] Rory Bunker and Teo Susnjak. 2022. The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research* 73 (2022), 1285–1322.

[3] Vitor Cerqueira, Luis Torgo, and Igor Mozetič. 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning* 109, 11 (2020), 1997–2028.

[4] Wei-Jen Chen, Mao-Jhen Jhou, Tian-Shyug Lee, and Chi-Jie Lu. 2021. Hybrid basketball game outcome prediction model by integrating data mining methods for the National Basketball Association. *Entropy* 23, 4 (2021), 477. doi:10.3390/e23040477

[5] F. Du and X. Yuan. 2020. A survey of visual analytics in sports. *IEEE Computer Graphics and Applications* (2020).

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (2018), 93. doi:10.1145/3236009

[7] Z. He and S.-B. Choi. 2025. Stacked ensemble learning for NBA game outcome prediction. *Expert Systems with Applications* (2025).

[8] T. Horvat, J. Job, R. Logozar, and Č. Livada. 2023. A data-driven machine learning algorithm for predicting the outcomes of NBA games. *Symmetry* 15, 4 (2023), 798.

[9] Lars Magnus Hvattum and Halvard Arntzen. 2010. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting* 26, 3 (2010), 460–470. doi:10.1016/j.ijforecast.2009.10.002

[10] S. Jain and H. Kaur. 2017. Machine learning approaches to predict basketball game outcome. In *Proceedings of the 3rd International Conference on Advances in Computing, Communication & Automation*. IEEE, 1–7.

[11] Reza Khanmohammadi, Sari Saba-Sadiya, Sina Esfandiarpour, Tuka Alhanai, and Mohammad Mahdi Ghassemi. 2024. MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs. *SN Computer Science* 5, 5 (2024), 628.

[12] P. H. Kvam and J. S. Sokol. 2006. A logistic regression/Markov chain model for NCAA basketball outcomes. *Naval Research Logistics* (2006).

[13] Chuxuan Li, Hao Zhang, Yuan Zhang, Jing Shen, and Ruopeng An. 2025. The application of artificial intelligence techniques in predicting game outcomes of professional basketball league: A systematic review. *PLOS ONE* 20, 6 (2025), e0326326.

[14] Weigeng Li, Neng Zhou, and Xiaodong Qu. 2024. Enhancing eye-tracking performance through multi-task learning transformer. In *International Conference on Human-Computer Interaction*. Springer, 31–46.

[15] Bernard Loeffelholz, Earl Bednar, and Kenneth W. Bauer. 2009. Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports* 5, 1 (2009), 7. doi:10.2202/1559-0410.1156

[16] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering* 31, 12 (2018), 2346–2363.

[17] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Vol. 30.

[18] Hans Manner. 2016. Modeling and forecasting the outcomes of NBA basketball games. *Journal of Quantitative Analysis in Sports* 12, 1 (2016), 31–41. doi:10.1515/jqas-2015-0088

[19] M. McFarlane. 2019. Evaluating win probability models in NBA end-game decision making. *Journal of Sports Analytics* (2019).

[20] Dragan Miljković, Ljubiša Gajić, Aleksandar Kovačević, and Zora Konjović. 2010. The use of data mining for basketball matches outcomes prediction. In *2010 8th International Symposium on Intelligent Systems and Informatics*. IEEE, 309–312. doi:10.1109/SISY.2010.5647440

[21] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[22] Nathan Koome Murungi, Michael Vinh Pham, Xufeng Dai, and Xiaodong Qu. 2023. Trends in machine learning and electroencephalogram (eeg): a review for undergraduate researchers. In *International Conference on Human-Computer Interaction*. Springer, 426–443.

[23] Nathan Koome Murungi, Michael Vinh Pham, Xufeng Caesar Dai, and Xiaodong Qu. 2023. Empowering computer science students in electroencephalography (eeg) analysis: A review of machine learning algorithms for eeg datasets. In *The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

[24] Y. Ouyang, X. Li, W. Zhou, W. Hong, W. Zheng, F. Qi, and L. Peng. 2024. Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology. *PLOS ONE* 19, 7 (2024), e0307478.

[25] Charles Perin, Romain Vuillemot, and Jean-Daniel Fekete. 2013. SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2506–2515. doi:10.1109/TVCG.2013.192

[26] J. Perricone, I. Shaw, and W. Swiechowicz. 2016. *Predicting results for professional basketball using NBA API data.* Technical Report. Stanford University. CS229 Technical Report.

[27] Xiaodong Qu, Matthew Key, Eric Luo, and Chuhui Qiu. 2024. Integrating HCI datasets in project-based machine learning courses: a college-level review and case study. In *International Conference on Human-Computer Interaction*. Springer, 124–143.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144. doi:10.1145/2939672.2939778

[29] J. A. Rodríguez. 2019. *"This game is in the fridge": Predicting NBA game outcomes.* Technical Report. Stanford University. CS229 Course Project.

[30] N. Sandholtz and L. Bornn. 2020. Markov decision processes for basketball shot selection. *Journal of Quantitative Analysis in Sports* (2020).

[31] Tse Saunders, Nawwaf Aleisa, Juliet Wield, Joshua Sherwood, and Xiaodong Qu. 2024. Optimizing the literature review process: Evaluating generative ai models on summarizing undergraduate data science research papers. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[32] E. Štrumbelj and I. Kononenko. 2012. A dynamic model for estimating win probabilities in basketball. *Journal of Quantitative Analysis in Sports* (2012).

[33] F. Thabtah, L. Zhang, and N. Abdelhamid. 2019. NBA game result prediction using feature analysis and machine learning. *Annals of Data Science* 6, 1 (2019), 103–116.

[34] J. van Haaren. 2021. Why trust your numbers? Explainable AI in sports analytics. *AI Magazine* (2021).

[35] M. Vračar, E. Štrumbelj, and I. Kononenko. 2016. Modeling basketball game outcomes using play-by-play data. *Expert Systems* (2016).

[36] J. Wang. 2023. Predictive analysis of NBA game outcomes through machine learning. In *Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence*. 46–55.

[37] Wensi Xie, Jingwen Dou, Yiran Wang, and Xiaodong Qu. 2025. From Theory to Play: A Review of EEG-Controlled Directional Games and Evaluation of Custom-Developed BCI Games. In *International Conference on Human-Computer Interaction*. 179–191.

[38] Isshin Yunoki, Guy Berreby, Nicholas D'Andrea, Yuhua Lu, and Xiaodong Qu. 2023. Exploring ai music generation: A review of deep learning algorithms and datasets for undergraduate researchers. In *International Conference on Human-Computer Interaction*. Springer, 102–116.

[39] K. Zhao, C. Du, and G. Tan. 2023. Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm. *Entropy* 25, 5 (2023), 765.

[40] K. Zhao, C. Du, and G. Tan. 2024. Basketball outcome prediction using graph attention networks with feature selection. *Neural Computing and Applications* (2024).

[41] Chenxu Zhu, Yiming Xu, and Xiaodong Qu. 2025. GViT: Combining Convolutional and Transformer Layers for Spatial-Temporal EEG Analysis. In *International Conference on Human-Computer Interaction*. Springer, 433–442.