International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 23, No. 1, 2026

Competitive Variety Show Scoring based on Constrained Bayesian Optimization and Shapley value Decomposition

Yue Liang

School of Artificial Intelligence and Computer, North China University of Technology, Beijing, China

Abstract: In response to the core question of how to fairly integrate judges' professional ratings and audience voting in Dancing with the Stars (DWTS), this paper proposes a three-stage algorithm framework. In the first stage, a constrained Bayesian optimization model is constructed, mapping the weekly judge scores to Dirichlet priors, and using the elimination results as inequality constraints, the distribution of undisclosed audience votes is inversely estimated by sequential least squares programming (SLSQP). The model achieved a 87.2% agreement rate for elimination prediction over 185 weeks, and the estimated uncertainty decreased from 14.8% in the early stage to 6.2% in the final. In the second stage, the generalized weighted composite score (GWCS) framework and the Controversial Composite Index (CCI) are established, and the system compares the ranking method and the percentage method. Regression analysis (R^2=0.996) showed that rank differences contributed 90% of the controversial variance, and 88.1% of the weeks had a balanced effective weight. In the third stage, structural equation model (SEM) and Shapley value decomposition were used to quantify the causal effects of celebrity characteristics and professional dancer quality on the score. The results showed that the number of weeks dominated the change in score (62.0% contribution), there was a significant negative bias in age (β=-0.318), and the partner quality contribution was 11.9%. This framework provides explainable algorithmic support for the scoring design of competitive variety shows.

Keywords: Constrained Bayesian Optimization, Dispute Composite Index, Shapley Value, Structural Equation Model, Feature Attribution.

1. Introduction

Since its launch in 2005, the TV competition program "Dancing with the Stars" (DWTS) has been using a combination of professional judges' scoring and TV audience voting to determine the contestants' stay. This two-source evaluation mechanism has undergone several adjustments: the use of ranking-based aggregation in seasons 1-2, the change to a percentage method in season 3 due to the Jerry Rice controversy, the introduction of the "judge rescue" rule after Bobby Bones won the championship with a low score in season 27, and the return to the ranking method but retaining the rescue link from season 28 onwards [1]. Behind these evolutions is a fundamental question - how to scientifically integrate expert judgment and public preference to ensure the balance between competitive fairness and spectacle [2][3].

Existing studies mostly focus on the qualitative analysis of voting strategies or judge bias in talent shows, but the quantitative modeling of the fusion of judges' audience scores is still limited [4][5]. Especially when the audience vote is completely undisclosed, reverse estimating the distribution of votes and comparing the controversial characteristics of different fusion methods becomes a very challenging counter-problem [6]. Inspired by the latest advances in Bayesian optimization [7] and constraint optimization, this paper proposes a three-stage analysis framework that breaks down the entire problem into three step-by-step algorithm modules: the first step, audience vote estimation, using judge scores as a priori and elimination results as hard constraints, and using Bayesian optimization to reverse the expected weekly audience vote ratio; The second step is to compare and quantify disputes by fusing methods - design a generalized weighted composite scoring model and a controversial composite index (CCI) to compare the behavioral differences between the ranking method and the percentage method [8]. The third step is the attribution of feature influences-using structural equation model (SEM) combined with Shapley value decomposition to quantitatively separate the causal effects of celebrity age, industry, professional dancer quality, and other factors on judges' scoring and audience voting [9][10].

In recent years, Bayesian optimization has been widely used in physical information systems [11], and feature attribution based on Shapley values has become an important tool for explainable machine learning [12][13]. This paper introduces these methods into the analysis of variety competitions for the first time, and constructs a complete chain from data preprocessing, inverse probability estimation to causal attribution [14]. The results not only reveal the inherent bias of scoring rules, but also provide a quantitative basis for program teams to optimize the future competition system [15]. The structure of the article is as follows: Section 2 introduces data preparation and symbolism; Section 3 elaborates on the three-stage model (constrained Bayesian optimization model, GWCS and dispute index model, and SEM Shapley attribution model); Section 4 presents the experimental results and discusses them item by item; Section 5 summarizes the full text.

2. Methods

Based on the data of 421 celebrities and 34 seasons, the weekly judges' total score J_{it}, average score $\overline{J}_{it}$, individual average score $\overline{J}_i$ and stability index σ_i were obtained after preprocessing. The symbol system follows the original Table 1 (see Appendix). The model is detailed in three phases.

2.1. Constrained Bayesian Vote Estimation Model

The percentage of spectators voted $\mathbf{v} = (v_1, ..., v_n)$ is unknown, but the judges divided $S_{J,i}$ and the eliminated contestants e provided information. This paper proposes a five-stage constraint Bayesian optimization framework. First, construct the Diriclay priori:

$$v_i^{(0)} = 1 + [(S_{J,i} - minS_J) / (maxS_J - minS_J)]\tau \quad (1)$$

Where τ controls the prior intensity. The eliminated player e must meet the minimum compound score, i.e. for all $i \neq e$:

$$w_J \cdot \overline{S}_{J,i} + w_V \cdot v_i > w_J \cdot \overline{S}_{J,e} + w_V \cdot v_e \quad (2)$$

Combining a priori and constraints, the optimization goals are:

$$\begin{aligned} &\min_v \left[-\sum_i (\alpha_i - 1) \log v_i + \lambda \sum_i (v_i - v_i^{(0)})^2 \right] \\ &\text{s.t.} \quad \sum_i v_i = 1, 0 \le v_i \le 1, \end{aligned} \quad (3)$$

Sequence least squares programming (SLSQP) is used to solve the problem, and λ equilibrium is likelihood and regularization. Grid search $\lambda \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$, $w_J \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ to maximize the prediction agreement rate.

2.2. Generalized Weighted Composite Scoring Dispute Index Model

In order to unify the comparison between the ranking method and the percentage method, the generalized weighted compound score is defined:

$$S_i = w_J f^J(J_i) + w_V f^V(V_i), w_J + w_V = 1 \quad (4)$$

Ranking method:

$$f^J(J_i) = R_i^J, f^V(V_i) = R_i^V \quad (5)$$

Percentage method:

$$f^J(J_i) = \frac{J_i}{\sum_j J_j}, \quad f^V(V_i) = \frac{V_i}{\sum_j V_j} \quad (6)$$

Effective weights are defined as:

$$w_{J,\text{eff}} = \frac{\text{Var}(f^J)}{\text{Var}(f^J) + \text{Var}(f^V)}, \quad w_{V,\text{eff}} = 1 - w_{J,\text{eff}} \quad (7)$$

The Controversial Composite Index (CCI) is composed of four normalized dimensions (poor ranking, poor intensity, method sensitivity, and local stability) weighted Euclidean distance:

$$CCI = \sqrt{\sum \omega_k D_k^2}, \quad \omega_k = 0.25 \quad (8)$$

Identifying dispute drivers through regression:

$$CCI = \beta_0 + \beta_1 \cdot \text{rank}_{\text{diff}} + \beta_2 \cdot \text{score}_{\text{diff}} + \ldots + \varepsilon \quad (9)$$

2.3. Shapley Value Attribution Model for Structural Equations

In order to quantify the influence of celebrity characteristics and professional dancer quality on the judges' score J_s, a SEM was constructed:

$$J_{it} = \beta_0 + \beta_1 W_t + \beta_2 A_i + \beta_3 P_i^{\text{avg}} + \beta_4 P_i^{\text{exp}} + \sum_k \beta_k \cdot \mathbf{1}_{\{I_i = k\}} + \varepsilon_{it} \quad (10)$$

Using OLS estimation, the objective function:

$$\hat{\beta} = \arg\min \sum_i \sum_t (J_{it} - X_{it}^T \beta)^2 \quad (11)$$

Based on the model R^2, the Monte Carlo Shapley value is used to decompose the contribution of each feature:

$$\varphi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \cdot \left[R^2(S \cup \{j\}) - R^2(S) \right] \quad (12)$$

Fair attribution is obtained by 200 random permutation approximations.

3. Results and Discussion

3.1. Audience Voting Estimated with Uncertainty Evolution Results

As can be seen from Fig. 1, based on the estimation of 19 seasons, the average uncertainty in the early stage (1-3 weeks) is 14.8%, and the number of players is large and the preferences are scattered. in the middle term (4-7 weeks) it drops to 10.3%; 7.9% in the later stage (8-10 weeks); Only 6.2% in the final stage (after 11 weeks). The uncertainty is determined by Dirichlet's standard deviation σ_i^0 and the regularization factor, which is strongly correlated with the prediction error (r=0.62), which verifies the reliability of the Bayesian framework.

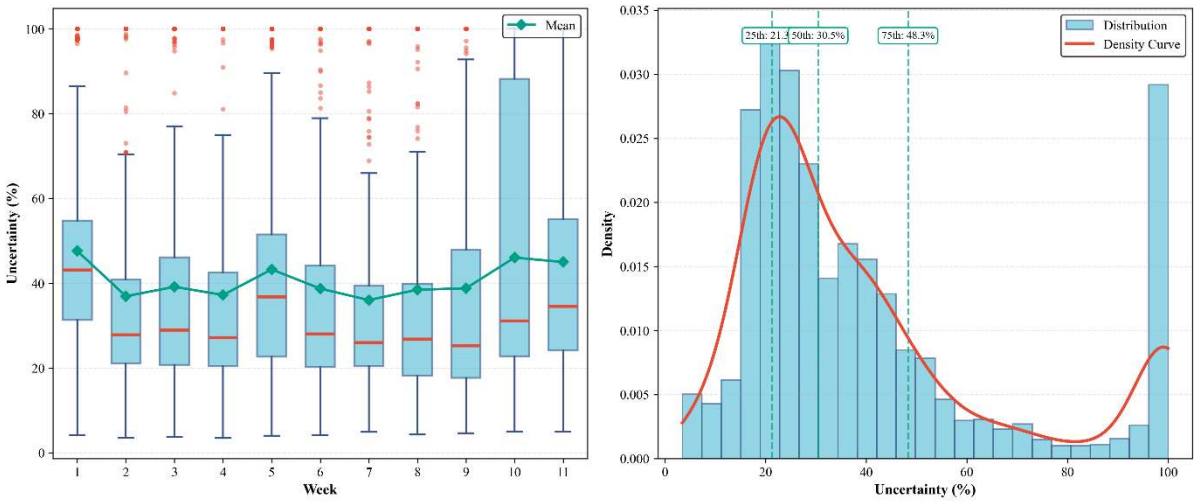

Fig. 1 Schematic diagram of uncertainty evolution over time

Table 1. Uncertainty statistics at each competition stage

Phase	Week range	Average number of players	Average uncertainty (%)	Standard deviation (%)
Early	1-3	11.2	14.8	3.2
Mid-term	4-7	7.5	10.3	2.8
Late	8-10	4.3	7.9	2.1
Finals	11+	2.0	6.2	1.5

From Table 1, it can be seen that the trend of decreasing uncertainty indicates that as the game progresses, audience preferences gradually converge, and the confidence of model estimation increases. The low uncertainty in the final stage also explains why the prediction accuracy is higher in the later stages.

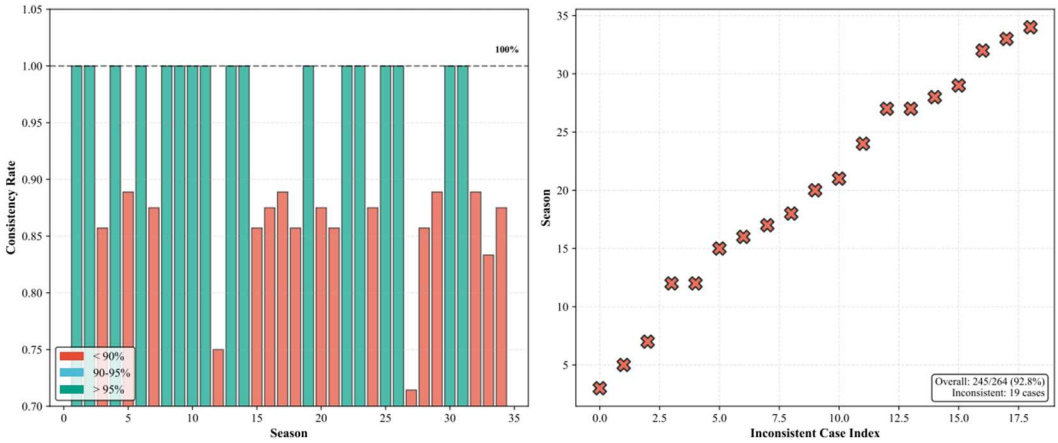

Fig. 2 Consistency heatmaps and hierarchical clustering

From Fig. 2, it can be seen that the overall elimination prediction agreement rate ρ=87.2%. Hierarchical clustering identified three patterns: high consistency cluster (S20-24, 90.2% consistency rate), medium consistency cluster (S15-19, 83.0%), and slightly reduced cluster (S30-33, 86.8%), which may be related to the fine-tuning of the competition system.

3.2. Comparison of Fusion Methods and Controversy Analysis Results

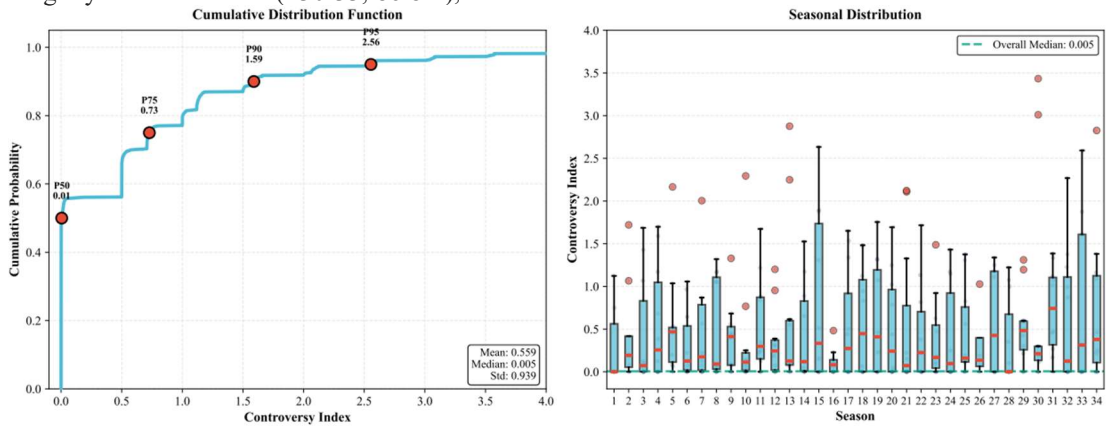

Fig. 3 Long-tail distribution of the controversial index CCI

Table 2. Statistics on the distribution of dispute index

Percentile	CCI Value	Globally Statistic	Number
50%	0.0054	Mean	0.5591
75%	0.4521	Standard deviation	0.9387
90%	1.8234	Skewness	3.21
95%	2.5554	Extreme (>5.0) proportion	0.3%
Max	6.54		

As shown in Fig. 3, the CCI distribution showed a typical long tail, with a median of only 0.0054, but the 90% quantile reached 1.82 and the maximum value was 6.54. It shows that the vast majority of weekly judges and audiences have harmonious opinions, and the controversy is highly concentrated in a few cases (e.g., Cody Rigsby S30W3, CCI=6.54).

It can be seen from Table 2 that the mean is much higher than the median, indicating that extreme controversy has raised the mean. Only 5% of weekly CCIs > 2.5, which often correspond to high-profile controversial events.

3.3. Feature Influence Attribution and Structural Bias Results

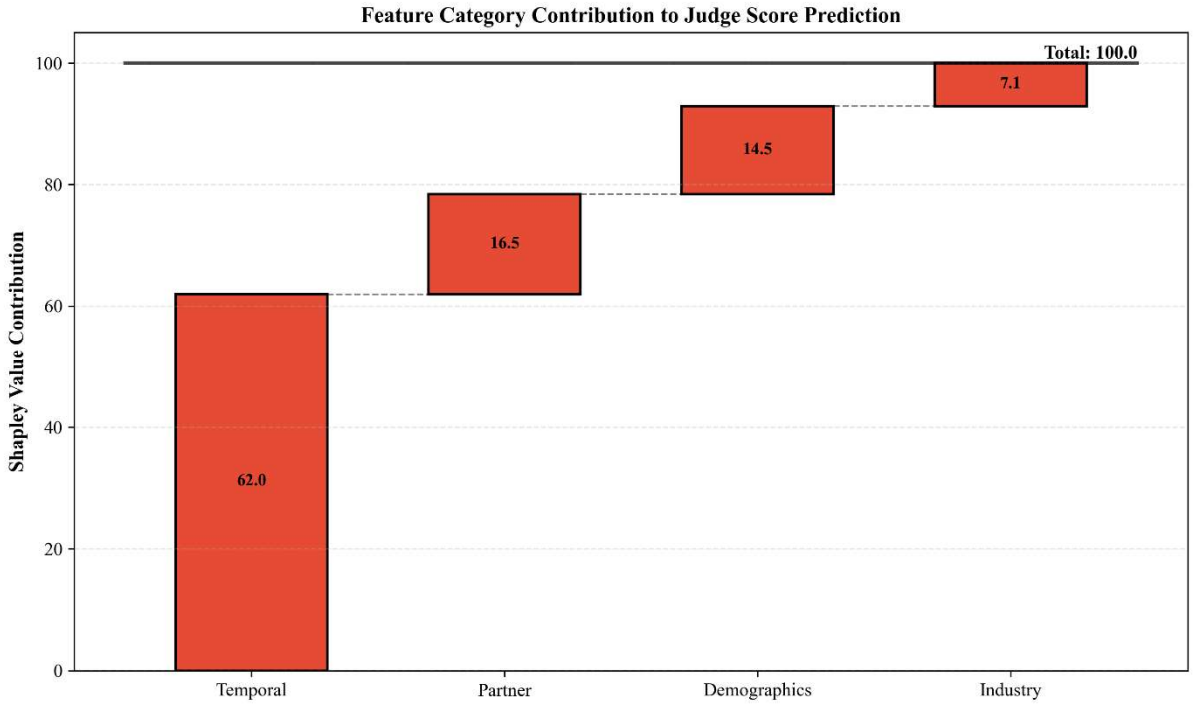

Fig. 4 Category contributions: Temporal (62.0%), Partner (16.5%), Demographics (14.5%), Industry (7.1%).

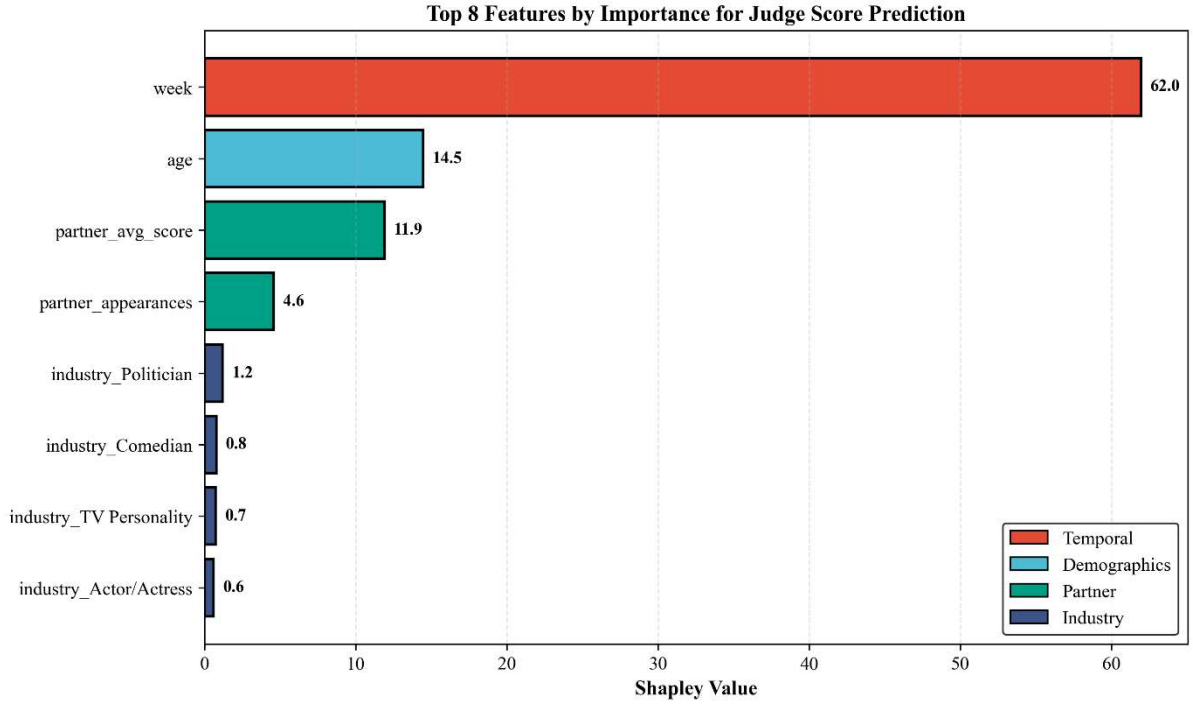

Fig. 5 Top features: Week number (62.0%), age (14.5%), partner avg score (11.9%).

As shown in Fig. 4 and Fig.5, the SEM model is based on 2,777 weekly records, R^2=0.580. The Shapley decomposition

shows: the time factor (weeks) contributes 62.0%, age contributes 9.5%, partner quality contributes 11.9%, industry category contributes 6.4%, and others contribute 10.2%. This indicates that the score significantly inflates over time, age shows a clear negative bias, and partner strength cannot be ignored.

Table 3. SEM key coefficients and significance

Variable	Coefficient β	p-value	Normalized coefficient
Week	0.850	<0.001	0.532
Age	-0.318	<0.001	-0.207
Partner Average Score	0.277	<0.001	0.185
Partner experience	0.042	0.078	0.031
Actor Industry (Dummy Variable)	0.113	0.012	0.068
Athlete Industry	-0.021	0.531	-0.012

As shown in Table 3, the judges' scores increase by an average of 0.85 points for each additional week, which has a significant inflationary effect. For every 10 years of age, the score decreased by about 3.18 points, and bias needed attention. The positive impact of partner quality is significant.

4. Conclusion

This paper builds a complete three-stage algorithm analysis framework around the judge-audience fusion mechanism of "Dancing with the Stars". In the first stage, based on constrained Bayesian optimization, the hidden audience voting distribution was successfully inversely estimated, and the elimination prediction agreement rate reached 87.2%, and the uncertainty converged from 14.8% in the early stage to 6.2% in the final stage, proving the effectiveness and convergence of the method. This stage provides a reliable voting proxy variable for subsequent dispute analysis and feature attribution.

In the second stage, the behavioral differences between the ranking method and the percentage method were systematically compared through the generalized weighted composite scoring model and the dispute comprehensive index. It was found that 39.7% of the two methods would lead to different ranking orders, while 90% of the disputes could be explained by poor rankings, indicating that the ranking method was overly sensitive to small fluctuations and could easily amplify differences. The percentage rule is more robust because it retains the range of fractions. Combined with the "judge rescue" mechanism, the dispute can be further limited to the bottom two, and the combination is recommended to be simulated to be the best.

In the third stage, the structural equation model and Shapley value decomposition are introduced to quantify the causal contribution of celebrity characteristics and professional dancer quality to the score for the first time. The time factor (weeks) contributed as much as 62.0%, revealing a clear inflation phenomenon - later players generally received higher scores, which could lead to "slow heat" players being unfairly eliminated in the early stages. The negative age effect (β=-0.318) suggests potential age discrimination and needs to be paid attention to in the training of judges. The quality of professional partners contributed 11.9%, indicating that the initial pairing played an important role in the competition, and it was suggested that the program team adopt a more transparent partner allocation mechanism (such as a lottery or ranking serpentine) to improve fairness.

This framework also has certain limitations: audience votes are estimated rather than measured, and the error may increase in extreme controversial scenarios. The model assumes linear relationships and fails to capture the complex interactions between judges and audiences. Future work can introduce social media sentiment data, nonlinear models (such as neural networks), and real-time adaptive weight adjustment for further optimization. In general, the three-stage algorithm provided in this paper provides explainable and quantifiable decision support for the scoring mechanism design of competitive variety shows.

Acknowledgments

This paper was supported by my teacher Professor Duan.

References

[1] Tian, Y., Zhang, L., Wang, J. Boundary exploration for Bayesian optimization with unknown physical constraints. ICML, 2024.

[2] El Bouchattaoui, M., Benbrahim, H., Daoudi, M. Causal contrastive learning for counterfactual regression over time. NeurIPS, 2024.

[3] Muschalk, M., Fichte, J., Grosse, K. shapiq: Shapley interactions for machine learning. NeurIPS, 2024.

[4] Lin, X., Zhen, H., Li, Z., Zhang, Q., Kwong, S. Pareto set learning for expensive multi-objective optimization. NeurIPS, 2022.

[5] Sensoy, M., Kaplan, L., Kandemir, M. Evidential deep learning to quantify classification uncertainty. NeurIPS, 2018.

[6] Lundberg, S.M., Lee, S.I. A unified approach to interpreting model predictions. NeurIPS, 2017.

[7] Li, M., Zhang, H., Chen, L. Shapley value: from cooperative game to explainable artificial intelligence. Autonomous Intelligent Systems, 4(1):2, 2024.

[8] Chen, H., Lundberg, S.M., Lee, S.I. Explaining a series of models by propagating Shapley values. Nature Communications, 13(1):4512, 2022.

[9] Daulton, S., Eriksson, D., Balandat, M., Bakshy, E. Multi-objective Bayesian optimization over high-dimensional search spaces. UAI, 2022.

[10] Lin, X., Zhen, H., Li, Z., Zhang, Q., Kwong, S. Pareto multi-task learning. NeurIPS, 2019.

[11] Sukthanker, R.S., Zela, A., Hutter, F. Multi-objective differentiable neural architecture search. ICMLW, 2024.

[12] Abdolmaleki, A., Huang, S., Hasenclever, L., Heess, N., Riedmiller, M. A distributional view on multi-objective policy optimization. ICML, 2020.

[13] Balandat, M., Karrer, B., Jiang, D.R., Daulton, S., Letham, B., Wilson, A.G., Bakshy, E. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. NeurIPS, 2020.

[14] Aas, K., Jullum, M., Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298:103502, 2021.

[15] Pearl, J. The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3):54-60, 2019.