Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 849–854 ISSN 2300-5963 ACSIS, Vol. 43
DOI: 10.15439/2025F5937

FedCSIS 2025 knowledgepit.ai Competition: Predicting Chess Puzzle Difficulty Part 2 & A Step Toward Uncertainty Contests

Jan Zyśko*, Michał Ślęzak†‡, Dominik Ślęzak*‡, Maciej Świechowski§‡
*Institute of Informatics, University of Warsaw, Poland
†Polish-Japanese Academy of Information Technology, Poland
‡QED Software, Poland
§Grail Team, Poland

***Abstract*—We summarize the results of the FedCSIS 2025 machine learning competition organized on the knowledgepit.ai online platform. We recall the competition's goals corresponding to estimations of the chess puzzle difficulty levels, we refer to the winning solutions, and we also compare the scope of this year's competition (and particularly the data available to competition participants) with its previous edition associated with the IEEE BigData 2024 conference. Finally, we discuss the new functionality of the knowledgepit.ai platform, which enables competition participants to submit additional uncertainty masks reflecting their assessment of test cases that are mostly problematic for their machine learning models.**

***Index Terms*—Human-Computer Interaction, Chess, Machine Learning Competitions, Uncertainty in Machine Learning**

I. Introduction

THE FedCSIS 2025 machine learning competition continues the successful blueprint established in the IEEE BigData 2024 edition [1], held at the knowledgepit.ai platform as well, addressing the challenge of predicting chess puzzle difficulty ratings from board configurations and solution sequences. While the core task remains unchanged – estimating Glicko-2 ratings that reflect human-perceived difficulty rather than engine-optimal evaluations – this second edition introduces significant enhancements in both data quality and modeling capabilities. The competition attracted 42 teams who collectively submitted 1185 solutions, with eight of the top ten teams contributing technical papers describing their approaches.

The primary innovations in this edition address key limitations of the previous competition. First, we integrated 22 precomputed success probability features generated by the Maia-2 model [2], providing participants with standardized human-aligned difficulty indicators across multiple rating bands and time controls. Second, we substantially improved test set quality through expanded data collection and a novel fairness algorithm that ensured balanced solving attempts across all puzzles, eliminating the rating convergence issues that plagued the first edition, where many puzzles remained clustered around the 1500 initialization value. These improvements resulted in more reliable ground truth labels and enabled more sophisticated modeling approaches, as evidenced by the winning solutions' creative use of ensemble methods, transformer architectures, and domain-specific feature engineering.

Beyond the core regression task, this competition pioneers a new dimension in machine learning competitions through the introduction of uncertainty masks – an extra challenge where participants identify the 10% of test cases their models find most problematic. This extension represents a step toward more interpretable and reliable machine learning systems, addressing the growing need for models that can communicate their (un)certainty levels alongside predictions.

The next sections describe the related work (Section II), the dataset (Section III), and the competition setup (Section IV). Further, we analyze the submitted solutions (Section V), discuss prediction errors and masking strategies of the top teams (Section VI), and conclude with implications for future competitions and research directions (Section VII).

II. Related Work

The challenges of the assessment of chess puzzle difficulty were discussed in our paper related to the first edition of this competition [1]. We also refer to [3]–[5] where the Top 3 solutions submitted to the first edition were reported.

The importance of machine learning competitions was also discussed in our previous papers [6]–[10]. In particular, we investigated possibilities of utilizing such competitions to build a platform for automatic evaluation of skills of data scientists [11]. As already mentioned, the new idea of extending the format of knowledgepit.ai competitions with uncertainty masks corresponds to that aspect of our research as well.

This new aspect of our competitions corresponds to an important topic in machine learning [12]. Although the most problematic cases for machine learning models are not necessarily the most "uncertain" ones, measures of uncertainty are often used, e.g., in active learning [13] and model diagnostics [14]. In our research, we estimate a given object's uncertainty using a neighborhood of cases having the same or similar values over subsets of attributes that are significant to the given model or decision problem [15]. The attribute selection mech-

anism described in [15] was also applied to build a benchmark machine learning model for this particular competition.

III. New Chess Puzzle Dataset

This competition builds upon the dataset foundation established in the previous IEEE BigData 2024 edition [1], while introducing a more robust test set and a significant enhancement to support more sophisticated modeling approaches.

A. Core Dataset Structure

Similarly to the previous edition, the dataset consists of chess puzzles with basic fields including PuzzleId, FEN (board position), Moves (puzzle solution), and the target Rating to be predicted. The training dataset retains the extra metadata fields from the original competition: RatingDeviation, Popularity, NbPlays, Themes, GameUrl, and OpeningTags.

B. Success Probability Features

The primary innovation in this edition is the addition of 22 success probability features generated using the Maia-2 model. Unlike the original Maia models that required separate trained models for each rating group [16], Maia-2 employs a unified architecture that jointly models success probabilities across all skill levels [2]. This model takes as input the chess position, player rating, and game type (rapid or blitz), outputting predicted success probabilities for solving the puzzle. Specifically, the Success_prob fields represent the estimated probability that a player of a given rating level and game type would correctly solve each puzzle. These 22 fields cover different rating ranges and both rapid and blitz time controls, providing participants with rich, model-generated features that capture human solving patterns across the skill spectrum.

Inspired by the innovative approaches observed in the top solutions from the previous competition [3]–[5], this enhancement aimed to provide participants with features that better capture the human element of puzzle-solving difficulty. Rather than requiring each team to independently extract chess engine evaluations for millions of puzzles, the Maia-2-generated success probabilities offer a standardized set of features that encode human-like assessment of puzzle difficulty across different skill levels.

C. Improved Test Set Quality

A significant improvement over the previous competition relates to the quality and distribution of puzzle attempt data. The IEEE BigData 2024 competition suffered from insufficient rating convergence in the test set, with many puzzles remaining near their initialized rating of 1500 due to limited solving attempts [1]. To address this limitation, the FedCSIS 2025 edition implemented two key enhancements:

- Expanded data collection: A larger pool of chess players was recruited to solve puzzles, resulting in substantially more solving attempts across the test set.
- Fairness algorithm: A novel algorithm was integrated into our lichess fork that prioritized presenting puzzles with fewer total attempts (regardless of correctness) to users. This ensured a more balanced distribution of solving attempts across all puzzles, leading to better rating convergence and more reliable ground truth labels.

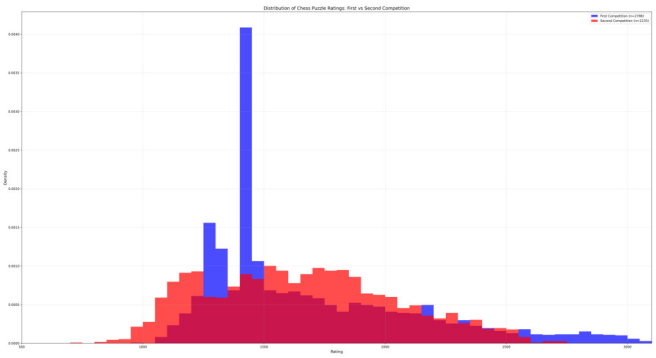

Fig. 1: Distribution of puzzle ratings in test sets: First competition (IEEE BigData 2024, blue) vs. Second competition (FedCSIS 2025, red). The first edition shows a pronounced spike at 1500 (the initialization value), while the second edition exhibits a more natural distribution with better rating convergence across the full spectrum.

Figure 1 illustrates the dramatic improvement in rating distribution quality. The first competition's test set (blue) exhibits a sharp peak around 1500—the default initialization value—indicating that many puzzles received insufficient solving attempts to converge to their true difficulty. In contrast, the FedCSIS 2025 test set (red) displays a smoother, more naturally distributed profile spanning from approximately 800 to 3000, demonstrating successful rating convergence. This improvement directly translates to more reliable ground truth labels for model evaluation.

IV. Competition Task

The core regression task of the FedCSIS 2025 Challenge remained fundamentally the same as in the first edition [1], requiring participants to predict difficulty ratings for chess puzzles based on their FEN board states and PGN solution sequences. The primary difference was the dataset size, which contained 2283 puzzles after removing those with fewer than 10 solving attempts during data annotation. As before, predictions were evaluated using Mean Squared Error (MSE) between predicted ratings $\hat{y}_i$ and ground-truth ratings y_i, computed as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{1}$$

where n denotes the number of test samples.

A novel extension to this edition introduced an uncertainty estimation task, whereby participants were invited to submit binary masks identifying the 10% of test puzzles for which their predictions were most likely to be erroneous. The mask $M \in \{0,1\}^N$ required each participant to flag exactly 10% of the test set as uncertain, with $M_i = 1$ indicating high uncertainty for puzzle i and the constraint $\sum_{i=1}^{N} M_i = N \times 0.1$. The

TABLE I: Top 10 submissions in the FedCSIS 2025 knowledgepit.ai competition. The evaluation metric was the mean squared error – the lower Final (and preliminary, Pre.) score, the better. The Subs column reflects the total number of submissions. The Mask column reports the uncertainty ratio – again, the lower, the better (the best score bolded). Eight out of Top 10 teams submitted conference papers describing their solutions. Nine teams participated in the additional uncertainty mask contest.

Rank	Team	Paper	Final	Pre.	Subs	Mask	Approach
1	ousou	[17]	**52.3k**	**55.4k**	89	1.64	Fine-tuned Maia-2 embeddings combined with other features through LightGBM, ensembled via CatBoost with linear scaling.
2	bread emoji	[18]	54.4k	58.2k	240	1.71	Two-stage approach with ensembled neural chessboard embedders, pre-RNN board pooling, and logarithmic rescaling.
3	transformer_enjoyer	[19]	55.9k	58.9k	117	**1.43**	Spatial transformer with masked-square reconstruction and solution policy prediction, rescaling borrowed from [3].
4	ToDoFindATeamName	–	57.5k	61.5k	136	–	–
5	Cyan	[20]	61.0k	67.1k	93	1.59	Multi-variant ensemble of Maia embeddings, solution sequences, and engine success probabilities with nonlinear calibration.
6	neuro	[21]	62.6k	66.5k	81	1.68	Four-stage pipeline with Elo-banded base models, gradient boosting, structural calibration via failure distributions, and ensemble averaging.
7	xyz	[22]	62.7k	66.6k	25	1.70	Three-stage pipeline with four Elo-banded MLP base models, stacking ensemble, and structure-aware residual correction.
8	DML	[23]	63.0k	66.7k	73	1.65	Three-stage pipeline combining GB models, multi-modal CNN with EfficientNetB3-rendered board images, and XGB residue stacking.
9	Ru	[24]	67.5k	70.4k	53	1.62	Three-stage pipeline with MobileNetV2 board images, LightGBM residual refinement, and adjustments based on failure distribution.
10	Feiwyth	–	68.1k	70.7k	17	1.58	–

rationale was that by replacing the predicted ratings at these masked positions with their ground-truths, the evaluation could assess a model's ability to identify its own confidence.

The uncertainty masks were evaluated using the Uncertainty Ratio metric, defined as:

$$\text{UR} = \frac{\mathcal{N}}{\mathcal{P}}, \tag{2}$$

where $\mathcal{N}$ represents the New Score (MSE after replacing masked predictions with ground truth) and $\mathcal{P}$ represents the Perfect Score (the minimum achievable MSE if the 10% highest-error samples were perfectly identified and masked). An optimal uncertainty ratio approaches 1.0, indicating perfect identification of the most erroneous predictions. For example, a ratio of 1.2 would indicate that the submitted mask captured errors that, when corrected, achieved 83.3% (1/1.2) of the theoretically optimal improvement, demonstrating reasonably good uncertainty estimation while leaving room for improvement in identifying the truly most difficult cases.

V. Competition Overview

Table I presents the top 10 submissions in the FedCSIS 2025 knowledgepit.ai competition, where teams competed to predict chess puzzle difficulty using mean squared error as the evaluation metric. Nine of the top 10 teams submitted conference papers describing their solutions and participated in the additional uncertainty mask contest.

We refer to Table I and papers [17]–[24] for detailed descriptions of particular competition solutions. In general, the best-performing models leveraged pretrained neural chess embeddings – particularly from the Maia family of models designed to mimic human play at various skill levels – combined with ensemble methods and gradient boosting techniques. The winning team achieved an MSE of 52.3k by fine-tuning Maia-2 embeddings and combining them with other features through LightGBM, followed by CatBoost ensembling with linear scaling. The second-place team (54.4k MSE) employed a two-stage approach with ensembled neural chessboard embedders and pre-RNN board pooling, while the third-place team (55.9k MSE) utilized spatial transformers with masked-square reconstruction. A common pattern across successful approaches was the use of multi-stage pipelines: initial feature extraction via neural board embeddings (Maia-1, Maia-2, or Leela), intermediate refinement through gradient boosting methods (LightGBM, XGBoost, or CatBoost), and final calibration via post-processing techniques addressing distributional shifts.

For uncertainty estimation, most teams developed mask prediction strategies that identified the least reliable predictions, achieving uncertainty ratios between 1.43 and 1.71. The transformer_enjoyer team achieved the best uncertainty ratio of 1.43 despite placing third overall, suggesting that model confidence estimation and raw prediction accuracy represent distinct challenges. Several approaches also incorporated the competition-provided Maia-2 success probabilities – precomputed estimates of puzzle completion likelihood for players at different skill levels – as auxiliary features, though their integration strategies varied considerably across teams.

VI. Error Analysis and Masking Strategies

A. Analysis of Prediction Errors

We conducted several post-hoc analyses of the rating prediction task. Figure 2 shows prediction errors by puzzle difficulty. A U-shaped pattern is visible: models perform worse on very easy (< 1000) and very hard (> 2400) puzzles, while medium-difficulty puzzles are predicted more accurately. This behavior is consistent with the MSE objective, which emphasizes outliers at both ends of the scale.

We then analyzed performance as a function of rating uncertainty (Figure 3). Prediction errors increase monotonically with the Glicko rating deviation of puzzles, indicating that puzzles with higher inherent rating uncertainty are systematically

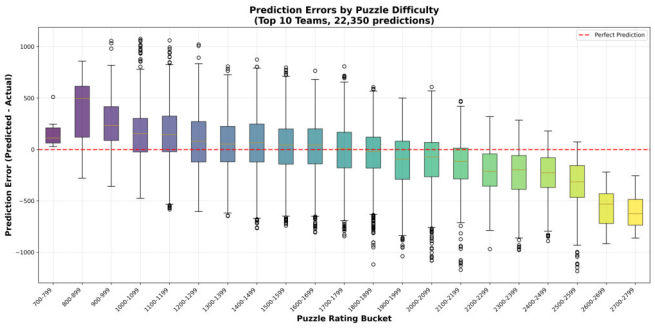

Fig. 2: Prediction errors by puzzle difficulty.

Fig. 3: Prediction errors by rating uncertainty.

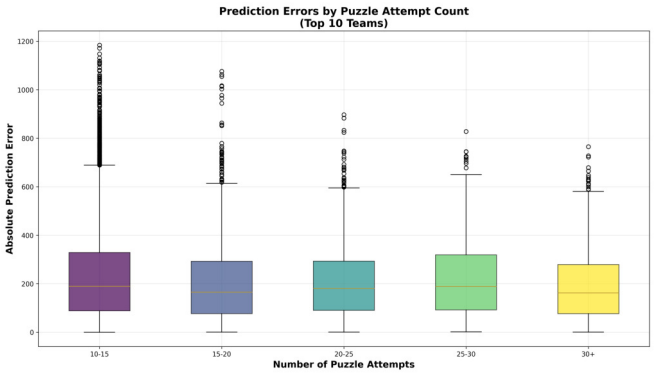

Fig. 4: Prediction errors by the number of attempts during data annotation phase.

harder to model. This suggests that rating deviation, rather than the number of attempts, is a more reliable signal for deciding how well a puzzle's difficulty has been established.

Finally, we investigated whether the number of puzzle attempts during tagging correlates with error (Figure 4). We found no significant trend: puzzles with 10–15 attempts yield error levels comparable to those with 30 or more attempts. Taken together, these findings imply that once a puzzle's Glicko deviation has stabilized, additional attempts provide little benefit. Thus, rating deviation should guide tagging efforts more directly than attempt count.

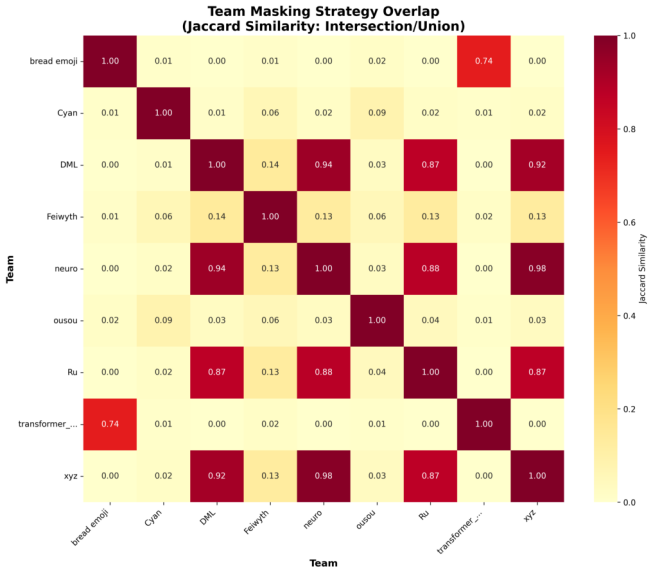

Fig. 5: Jaccard similarity matrix showing overlap between teams' masking strategies.

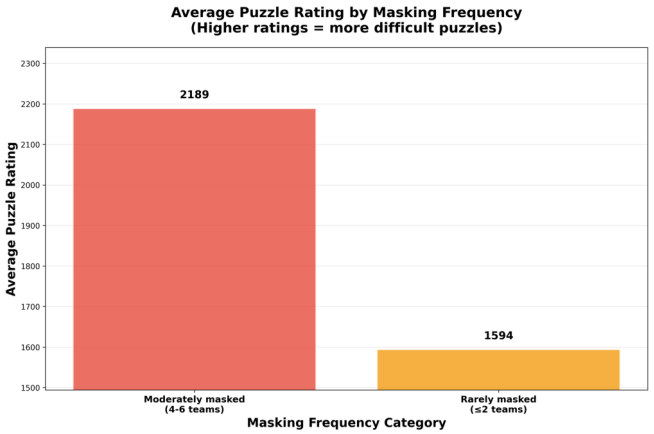

Fig. 6: Puzzle characteristics grouped by masking frequency across teams.

B. *Masking Strategy Analysis*

We next examined the masking strategies employed by teams, focusing both on the convergence of approaches and the characteristics of commonly excluded puzzles. Figure 5 shows the Jaccard similarity between teams' masking choices. Several teams converged on remarkably similar strategies, with DML, neuro, Ru, and xyz achieving similarity scores of 0.87–0.98. This suggests that these teams independently identified a common set of problematic puzzles. By contrast, teams such as bread emoji and transformer_enjoyer pursued more distinctive approaches, showing consistently lower overlap with others.

Beyond team overlap, we analyzed which puzzles were most frequently masked (Figure 6). A clear trend emerges: puzzles targeted by many teams tend to be significantly more difficult ($\approx$ 2200) than those rarely masked ($\approx$ 1600). This indicates

that masking was not random but systematically biased toward high-difficulty puzzles, which teams collectively judged as less reliable or harder to model.

VII. Conclusions

The FedCSIS 2025 knowledgepit.ai competition highlighted the effectiveness of combining neural chess embeddings with ensemble methods and calibration strategies for predicting puzzle difficulty. The top-performing teams consistently employed multi-stage pipelines in which pretrained embeddings such as Maia-2 served as the foundation for downstream boosting models and refined scaling. Despite differences in architecture, a recurring theme was the integration of multiple feature sources – ranging from engine success probabilities to solution policy predictions – followed by systematic post-processing. The additional mask contest further demonstrated that raw predictive accuracy and confidence estimation are distinct goals – while the winning team achieved the lowest MSE overall, the strongest uncertainty modeling was achieved by a different team. This emphasizes the importance of addressing not only point prediction but also model reliability.

Post-hoc analyses of prediction errors revealed consistent patterns across submissions. Errors followed a U-shaped trend with respect to puzzle difficulty, increasing at both extremes of the scale, while rating deviation emerged as a more reliable indicator of prediction uncertainty than attempt count. These findings suggest that puzzle ambiguity, as quantified by Glicko deviation, poses a fundamental barrier to modeling accuracy and that additional attempts beyond stabilization yield limited benefit. Overall, the competition illustrates the dual challenges of modeling puzzle difficulty and quantifying uncertainty, pointing to promising directions for future research in hybrid modeling and calibrated confidence estimation.

In the future, we intend to embed the mechanism of uncertainty masks into all challenges held at the knowledgepit.ai platform – not only as additional tasks but as the main competition objectives. This is because the ability to identify the hardest cases in the data is the key element of advanced machine learning processes [13], and therefore, we need to develop tools for measuring whether the competition participants possess the appropriate skills in this regard [11]. While doing this, however, we need to remember that uncertainty measures are not necessarily the only way to identify those most problematic cases for machine learning models.

As another future topic, encouraged by the aforementioned fact that most of the successful teams relied heavily on non-trivial feature engineering, we are going to further improve our benchmark competition model, this time taking into account also the attributes gathered from papers [17]–[24], organizing them within semantically meaningful groups, and deriving ensembles of their subsets as proposed in [15].

Acknowledgments

This research was co-funded by Smart Growth Operational Programme 2014-2020, financed by European Regional Development Fund, in frame of project POIR.01.01.01-00-1070/21, operated by National Centre for Research and Development in Poland. We are also grateful to chess players who helped us with the difficulty ratings of puzzles in the test data.

References

[1] J. Zyśko, M. Świechowski, S. Stawicki, K. Jagieła, A. Janusz, and D. Ślęzak, "IEEE Big Data Cup 2024 Report: Predicting Chess Puzzle Difficulty at KnowledgePit.ai," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8423–8429. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10825289

[2] Z. Tang, D. Jiao, R. McIlroy-Young, J. M. Kleinberg, S. Sen, and A. Anderson, "Maia-2: A Unified Model for Human-AI Alignment in Chess," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10-15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024. [Online]. Available: http://papers.nips.cc/paper_files/paper/2024/hash/250190819ff1dda47cd23cecc0c5a69b-Abstract-Conference.html

[3] T. Woodruff, O. Filatov, and M. Cognetta, "The bread emoji Team's Submission to the IEEE BigData 2024 Cup: Predicting Chess Puzzle Difficulty Challenge," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8415–8422. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10826037

[4] A. Schütt, T. Huber, and E. André, "Estimating Chess Puzzle Difficulty Without Past Game Records Using a Human Problem-Solving Inspired Neural Network Architecture," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8396–8402. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10826087

[5] S. Björkqvist, "Estimating the Puzzlingness of Chess Puzzles," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8370–8376. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10825991

[6] A. Janusz, M. Przyborowski, P. Biczyk, and D. Ślęzak, "Network Device Workload Prediction: A Data Mining Challenge at Knowledge Pit," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 21, 2020, pp. 77–80. [Online]. Available: https://doi.org/10.15439/2020F159

[7] A. Janusz and D. Ślęzak, "KnowledgePit Meets BrightBox: A Step Toward Insightful Investigation of the Results of Data Science Competitions," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022, Sofia, Bulgaria, September 4-7, 2022*, ser. Annals of Computer Science and Information Systems, vol. 30, 2022, pp. 393–398. [Online]. Available: https://doi.org/10.15439/2022F309

[8] M. Wnuk, J. Dziuba, A. Janusz, and D. Ślęzak, "IEEE BigData Cup 2023 Report: Object Recognition with Muon Tomography Using Cosmic Rays," in *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, J. He, T. Palpanas, X. Hu, A. Cuzzocrea, D. Dou, D. Ślęzak, W. Wang, A. Gruca, J. C. Lin, and R. Agrawal, Eds. IEEE, 2023, pp. 6084–6091. [Online]. Available: https://doi.org/10.1109/BigData59044.2023.10386564

[9] A. Janusz and D. Ślęzak, "Predicting Frags in Tactic Games at knowledgepit.ai: ICME 2023 Grand Challenge Report," in *IEEE International Conference on Multimedia and Expo Workshops, ICMEW Workshops 2023, Brisbane, Australia, July 10-14, 2023*. IEEE, 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICMEW59549.2023.00006

[10] A. M. Rakicevic, P. D. Milosevic, I. T. Dragovic, A. M. Poledica, M. M. Zukanovic, A. Janusz, and D. Ślęzak, "Predicting Stock Trends Using Common Financial Indicators: A Summary of FedCSIS 2024 Data Science Challenge Held on knowledgepit.ai Platform," in *Proceedings of the 19th Conference on Computer Science and Intelligence Systems, FedCSIS 2024, Belgrade, Serbia, September 8-11, 2024*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 39, 2024, pp. 731–737. [Online]. Available: https://doi.org/10.15439/2024F7912

[11] D. Ślęzak, A. Janusz, M. Świechowski, A. Chądzyńska-Krasowska, and J. Kamiński, "Do Data Scientists Dream About Their Skills' Assessment? – Transforming a Competition Platform Into an Assessment Platform," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, Eds. IEEE, 2024, pp. 8403–8414. [Online]. Available: https://doi.org/10.1109/BigData62323.2024.10825378

[12] E. Hüllermeier and W. Waegeman, "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021. [Online]. Available: https://doi.org/10.1007/s10994-021-05946-3

[13] D. Kałuża, A. Janusz, and D. Ślęzak, "Evidence-theoretical Modeling of Uncertainty Induced by Posterior Probability Distributions," *International Journal of Applied Mathematics and Computer Science*, vol. 35, no. 1, 2025. [Online]. Available: https://doi.org/10.61822/amcs-2025-0003

[14] A. Janusz, A. Zalewska, Ł. Wawrowski, P. Biczyk, J. Ludziejewski, M. Sikora, and D. Ślęzak, "BrightBox – A Rough Set based Technology for Diagnosing Mistakes of Machine Learning Models," *Applied Soft Computing*, vol. 141, p. 110285, 2023. [Online]. Available: https://doi.org/10.1016/j.asoc.2023.110285

[15] A. Janusz, D. Kałuża, D. Ślęzak, and S. Stawicki, "Automatic Generation of Attributes based on Semantic Categorization of Large Datasets in Artificial Intelligence Models and Applications," *US Patent Application US-20250005436-A1*, 2025.

[16] R. McIlroy-Young, S. Sen, J. M. Kleinberg, and A. Anderson, "Aligning Superhuman AI with Human Behavior: Chess as a Model System," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 1677–1687. [Online]. Available: https://doi.org/10.1145/3394486.3403219

[17] S. Björkqvist, "Estimating the Difficulty of Chess Puzzles by Combining Fine-Tuned Maia-2 with Hand-Crafted and Engine Features," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F6497

[18] T. Woodruff, L. Imbing, and M. Cognetta, "The bread emoji Team's Submission to the 2025 FedCSIS Predicting Chess Puzzle Difficulty Challenge," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F6771

[19] S. Miłosz, "Pretraining Transformers for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F7603

[20] H. Xiao, D. Yu, X. Wen, L. Chen, and K. Fu, "Multi-Source Feature Fusion and Neural Embedding for Predicting Chess Puzzle Difficulty," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F2456

[21] L. Cen, J. Cen, and M. Song, "A Multi-Stage Framework for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F4532

[22] A. Liang, C. Liu, K. Wang, and E. Liu, "A Stacking-Based Ensemble Approach for Predicting Chess Puzzle Difficulty," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F1698

[23] M. Liu, J. Wang, Y. Hu, and D. Lin, "Hybrid Boosting and Multi-Modal Fusion for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F3675

[24] J. Chen, C. Liu, and Y. Gao, "Multi-Modal Deep Learning with Residual and Structure-Guided Refinement for Chess Puzzle Difficulty Prediction," in *Proceedings of the 20th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Bolanowski, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 43. Polish Information Processing Society, 2025. [Online]. Available: http://dx.doi.org/10.15439/2025F3227