Bibliotecas da Universidade Federal da Fronteira Sul - UFFS

Bonelli, Djonatan Riquelme Clein
Tree-Based Learning for Game Outcome Prediction / Djonatan Riquelme Clein Bonelli, João Luís Almeida Santos, Eduardo Vinicius Perissinotto Fiorentin, Andrei Carlesso Camilotto, Felipe Grando. -- 2025.
19 f.:il.

Orientador: Dr. Felipe Grando

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal da Fronteira Sul, Curso de Bacharelado em Ciência da Computação, Chapecó,SC, 2025.

1. Game outcome prediction. 2. Tree-based models. 3. Multi-agent systems. 4. Model interpretability. 5. Ensemble methods. I. Santos, João Luís Almeida II. Fiorentin, Eduardo Vinicius Perissinotto III. Camilotto, Andrei Carlesso IV. Grando, Felipe V. Grando, Felipe, orient. VI. Universidade Federal da Fronteira Sul. VII. Título.

Elaborada pelo sistema de Geração Automática de Ficha de Identificação da Obra pela UFFS com os dados fornecidos pelo(a) autor(a).

DJONATAN RIQUELME CLEIN BONELLI

APRENDIZADO BASEADO EM ÁRVORES PARA PREVISÃO DE RESULTADOS DE JOGOS

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação da Universidade Federal da Fronteira Sul (UFFS), como requisito para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Felipe Grando

Este Trabalho de Conclusão de Curso foi avaliado e aprovado pela banca avaliadora em: 25/11/2025.

BANCA AVALIADORA

Documento assinado digitalmente
gov.br
FELIPE GRANDO
Data: 27/11/2025 11:51:49-0300
Verifique em https://validar.iti.gov.br

Felipe Grando - UFFS

Documento assinado digitalmente
gov.br
CLAUNIR PAVAN
Data: 27/11/2025 15:28:47-0300
Verifique em https://validar.iti.gov.br

Claunir Pavan - UFFS

Documento assinado digitalmente
gov.br
MARCO AURELIO SPOHN
Data: 27/11/2025 13:31:16-0300
Verifique em https://validar.iti.gov.br

Marco Aurélio Spohn - UFFS

Tree-Based Learning for Game Outcome Prediction

Abstract

Predicting outcomes in complex multi-agent games is challenging due to imperfect information, stochastic events, and strategic interactions. We investigate interpretable tree-based models for outcome prediction in *Citadels*, a strategic board game that serves as a testbed for multi-agent dynamics. Using Optuna for hyperparameter optimization, we configure Decision Trees and Random Forests, with the latter consistently outperforming single trees. Prediction accuracy exceeds 60% in early game rounds and surpasses 90% in later rounds, while metric-specific optimization highlights trade-offs among precision, recall, and F1 score. showing that combinations of a few features can yield strong predictive signals when interaction effects are considered. Our results demonstrate that interpretable tree-based models can combine robust predictive performance with transparent explanations, offering insights into strategic behavior and informing the broader study of decision-making in complex multi-agent environments.

Keywords: Game outcome prediction, Tree-based models, Ensemble methods, Multi-agent systems, Strategic board games, Model interpretability

Artificial intelligence (AI) has long used games as testing grounds, where strategic depth, uncertainty, and multi-agent dynamics unfold in controlled environments [1,2]. From Deep Blue's *Chess* victory [3] to reinforcement learning breakthroughs in *Go* [4] and *Chess* [5], AI and machine learning have achieved superhuman performance, often through complex black-box models. Yet, as models grow in sophistication, the gap between accuracy and interpretability becomes increasingly critical, particularly in domains where transparency and actionable insights matter as much as raw performance [6,7].

Recent advances in stochastic and competitive games such as *Poker* [8], *Hearthstone* [9], *Dota 2* [10], and *League of Legends* [11,12] show how predictive models can anticipate victory and measure advantage, even with thousands of matches and complex interactions. However, most studies focus on large-scale eSports or classic games, leaving structured, turn-based board games underexplored. Unlike these domains, board games such as *Citadels* combine strategic depth, stochasticity, hidden information, and multi-agent interactions within a compact, well-structured environment. Players must balance district construction, role selection, and resource management, while adapting to opponents' actions and random draws. Victory depends on both

long-term planning and tactical responses, making *Citadels* an ideal testbed for interpretable models that capture strategic dynamics, relational features, and the effects of chance.

In this study, we address this gap through a case study on *Citadels*. We ask: (1) can tree-based models reliably predict match outcomes during the course of a game match? (2) which configurations and hyperparameters yield the best trade-off between accuracy and interpretability? And (3) to what extent can these models reveal the features and interactions that drive success or failure, thereby supporting both human decision-making and the design of adaptive AI agents?

Our results demonstrate that tree-based models can indeed predict match outcomes in *Citadels* with high reliability, achieving over 60% accuracy in early game stages and exceeding 90% in later stages. Random Forest models consistently outperformed single decision trees, confirming the advantage of ensemble architectures while preserving a high degree of interpretability. Hyperparameter optimization revealed clear trade-offs: deeper trees and larger ensembles improved accuracy (up to 0.78 overall) but reduced transparency, whereas shallower, precision-oriented models provided more compact and explainable decision structures. Feature-importance and SHAP analysis further indicated that no single variable dominates the prediction process; instead, multiple features contribute modestly, and their interactions, particularly between player progress and opponent performance, jointly explain most of the predictive variance. These findings collectively indicate that interpretable tree-based models can not only forecast outcomes with strong consistency but also elucidate the strategic dynamics driving victory in complex, multi-agent environments.

To detail these answers, we combine optimized tree-based models with global and local interpretability techniques to analyze how computational and strategic factors evolve throughout gameplay. In the following section, we present empirical results, feature-importance analysis, and stage-wise evaluations, highlighting how predictive accuracy can be reconciled with transparent explanations. Our findings contribute to interpretable machine intelligence in complex environments, bridging predictive performance with explainable insights.

1 Results

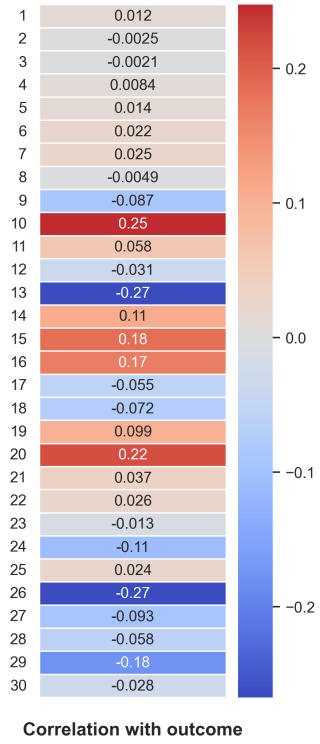
We report the main empirical results, spanning data characterization, model optimization, predictive performance, and interpretability. The synthetically generated dataset is first analyzed to uncover key feature distributions and correlations that shape model learning. Hyperparameter optimization with Optuna identifies configurations that balance performance and interpretability. The resulting Decision Tree and Random Forest models are then evaluated across accuracy, precision, recall, and F1 score, both globally and across individual game phases. Finally, global feature importance and SHAP-based interpretability analyzes reveal the dominant features and their interactions, offering insight into the strategic dynamics learned by the models.

Agent	Win Rate (%)	Avg. Score
SA1	36.97	21.87
SA2	32.12	19.64
SA3	32.12	21.42
SA4	21.21	17.51
SA5	7.27	12.30
SA6	6.67	15.87
Random	3.64	11.71

(a)

ID	Cat.	Description
1	GB	Round number
2	GB	Maximum built districts (global)
3–7	GB	Player partial scores (P1–P5)
8	AP	Gold amount
9	AP	No. of cards in hand
10	AP	No. of districts built
11	AP	Total cost of built districts
12	AP	Total cost of hand districts
13	AP	No. of district types built
14	AP	No. of district types in hand
15	AP	No. of low-cost built districts
16	AP	No. of high-cost built districts
17	AP	No. of low-cost districts in hand
18	AP	No. of high-cost districts in hand
19	AP	No. of special districts in hand
20	AP	No. of special districts built
21	AP	Rank of selected character
22–30	MVP	Same metrics for MVP (excluding private "in hand" features)

(b)



(c)

Fig. 1: Data characterization. (a) Performance of Specialist (SA1–SA6) and Random agents in the dataset. (b) The 30 input features capture both game board (GB) conditions and player states from the Active Player's (AP) perspective, combining the AP's internal state with observable information from the player with the highest partial score (MVP). (c) The heatmap shows Pearson correlation coefficients between each game-state variable and the match outcome (victory = 1, defeat = 0).

1.1 Data Characterization

We generated a synthetic dataset of game states representing interactions among agents derived from human-inspired strategies, including a stochastic Random agent. A full round-robin tournament was conducted with seven agents taken five at a time, yielding 11,550 games and 98,517 partial game-state samples, of which 38,082 (38.6%) corresponded to win labeled samples. Specialist agents (SA1–SA6) consistently outperformed the Random agent (Fig. 1a), though their differing play styles introduced desirable variability into the dataset, enriching the strategic diversity captured by the models.

To better understand the structure of this dataset and the relationships shaping predictive learning, we next examined feature dependencies and correlations. Given the

stochastic and multifactorial nature of *Citadels*, correlation analyzes were conducted to highlight the most informative variables and guide subsequent model regularization. The selected variables captured both the overall game board configuration (GB) and player-specific dynamics. Namely, those of the active player (AP) and the player with the highest partial score (MVP), as detailed in Fig. 1b and described in the Methods section (3.3).

Feature-label correlations suggest that no single feature is strongly predictive of victory in isolation (Fig. 1c). Instead, several district-related variables show moderate associations, collectively shaping outcome prediction. The most notable positive correlations are observed for the active player (AP): the number of districts built (+0.25), the number of special districts built (+0.22), the number of low-cost districts built (+0.18), and the number of high-cost districts built (+0.17). Conversely, negative correlations appear for the number of district types built by the MVP (-0.27), the number of district types built by the AP (-0.27), and the number of special districts built by the MVP (-0.18). These patterns suggest that victory emerges from a combination of structural and strategic building behaviors, where multiple complementary features jointly contribute to predictive performance.

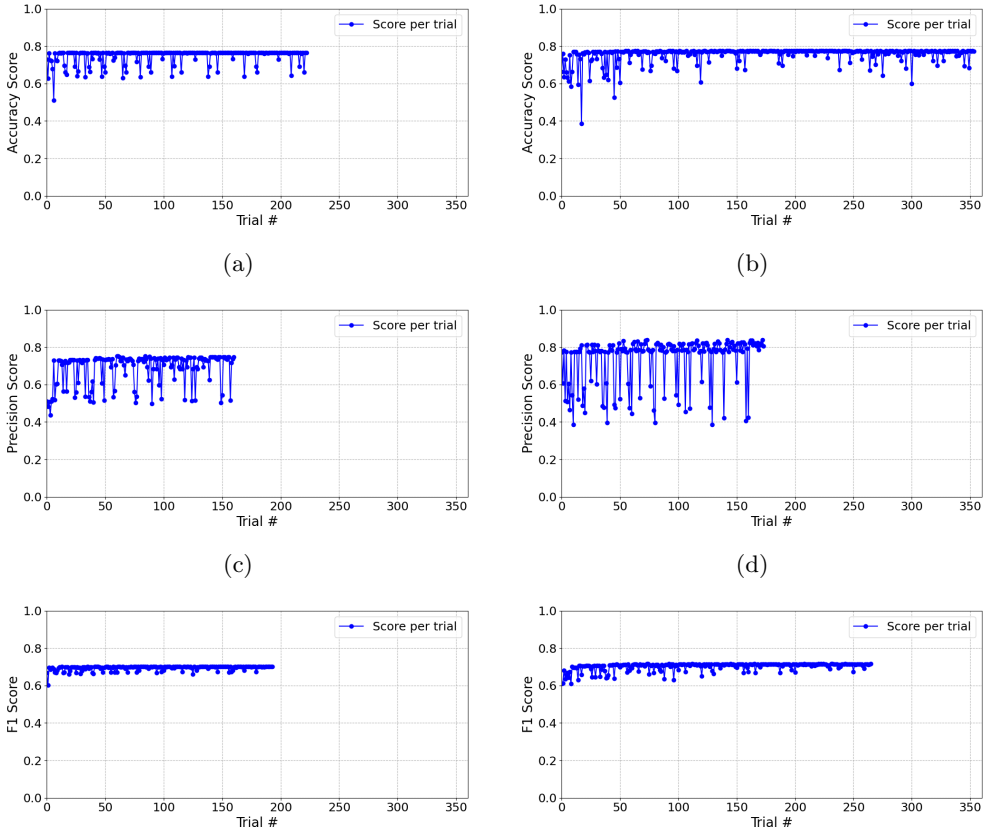
While informative, the feature space was not sufficiently redundant to justify dimensionality reduction, and all features were retained for training. These analyzes delineate the key structural relationships within the dataset, establishing the foundation for the model optimization procedures detailed in the following section.

1.2 Model Optimization

Next, we report the results of hyperparameter optimization and summarize key findings from the best-performing configurations. Models were trained through automated searches using Optuna. For clarity, model identifiers indicate the algorithm and the target metric: decision tree classifiers (CART) and random forests (RF) were optimized for accuracy (A), precision (P), or F1 score (F1). The implementation details of the Optuna optimization procedure are described in the Methods section (3.4).

Trial histories (Fig. 2 a–f) reveal distinct convergence behaviors. CART-P (Fig. 2c) required the fewest trials, stabilizing quickly after early fluctuations, whereas RF-A (Fig. 2b) demanded more iterations but rapidly discovered multiple near-optimal regions. On average, optimization required about 228 trials per study (191.7 for CART and 263.7 for RF). These results indicate that Random Forests, while generally achieving superior predictive power, require more nuanced hyperparameter adjustments to balance competing objectives effectively.

Fig. 2g summarizes the best-performing hyperparameters identified through Optuna optimization. Precision-optimized models converged to shallower trees, with low maximum depth and, in the case of Random Forests, a reduced number of estimators, reflecting their focus on selective and high-confidence decision boundaries. In contrast, F1-optimized models explicitly addressed class imbalance through adjusted class weights, achieving a more balanced trade-off between false positives and false negatives. Accuracy- and F1-oriented configurations generally favored deeper trees, often with no explicit constraint on maximum depth, while mitigating overfitting through higher min_samples_leaf and min_samples_split thresholds or by limiting the number



(a) (b) (c) (d) (e) (f)

Hyperparameter	CART-A	CART-F1	CART-P	RF-A	RF-F1	RF-P
max_depth	None	None	6	None	26	2
n_estimators	–	–	–	774	220	51
criterion	gini	gini	log_loss	log_loss	gini	entropy
min_samples_leaf	87	45	176–179	2	2	301
min_samples_split	466	264	143–167	15	46	227
class_weight	{0:1,1:1}	{0:1,1:2}	{0:1,1:1}	{0:1,1:1}	{0:1,1:2}	{0:1,1:1}

(g)

Fig. 2: Overview of the hyperparameter optimization convergence history and resulting configurations. (a) CART-A (accuracy-optimized Decision Tree), (b) RF-A (accuracy-optimized Random Forest), (c) CART-P (precision-optimized Decision Tree), (d) RF-P (precision-optimized Random Forest), (e) CART-F1 (F1-optimized Decision Tree), (f) RF-F1 (F1-optimized Random Forest), and (g) optimized hyperparameters for CART and RF models.

of estimators. No consistent pattern was observed regarding the chosen split criterion. Overall, Optuna-based tuning effectively uncovered distinct yet stable regions of optimal configurations, balancing model complexity and generalization capacity. These findings demonstrate how different optimization objectives shape the structural biases of decision-tree–based models toward complementary forms of predictive specialization.

Building on these optimized settings, the following step is to evaluate the predictive performance of each model.

1.3 Predictive Performance

Fig. 3a summarizes the performance of the optimized models across standard metrics (accuracy, precision, recall, and F1 score). As expected, models tended to perform best on the metric used for their optimization objective. The RF-A achieved the highest overall accuracy (0.7760), closely followed by CART-A (0.7648). Precision-oriented models, particularly RF-P, reached high precision (0.8383) but exhibited low recall (0.2609). F1-oriented models showed more balanced performance, with CART-F1 attaining the highest recall (0.8114) and RF-F1 achieving the highest F1 score (0.7160).

Several key insights emerge from these results. Precision-optimized models excel at identifying rare victories but exhibit low recall, reflecting trade-offs driven by class imbalance. F1-optimized models balance precision and recall, whereas accuracy-optimized models deliver strong overall performance without overemphasizing minority outcomes.

To examine temporal robustness, models were further evaluated across specified game intervals using a test set composed of 39.418 synthetic samples (Fig. 3 b–e).

Overall, model performance improved consistently as the game advanced, while the relative ranking among models remained largely stable. Accuracy exceeded 0.6 in the early phases and surpassed 0.85 by the final intervals. F1 scores (excluding RF-P) started above 0.4 and approached 0.9 in the later stages. Precision remained high for targeted models (0.85-0.9 at best), while recall exhibited greater variability early in the game (0.1-0.7) but converged near 0.9, except for RF-P, which consistently underperformed due to its precision-focused optimization.

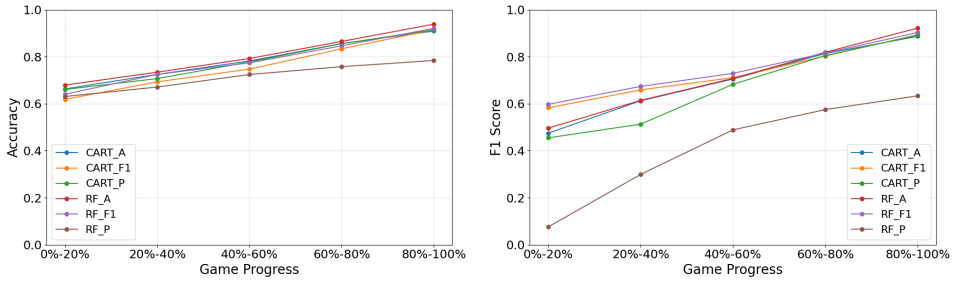
Metric-specific performance trends across models were observed as follows:

- **Accuracy** (Fig. 3b): RF-A consistently outperformed its CART counterpart across all phases, demonstrating the robustness of ensemble models.
- **F1 score** (Fig. 3c): RF-F1 led in the early stages, but RF-A surpassed it later, while CART-F1 remained stable yet slightly lower, reflecting metric-specific trade-offs.
- **Precision** (Fig. 3d): RF-P maintained strong precision in the early stages but gained little in the later stages; CART-P showed smaller, stage-dependent improvements, suggesting that simpler models retain relevance beyond their target metric.
- **Recall** (Fig. 3e): RF-P underperformed throughout, while CART-P achieved more consistent recall, although the highest values occurred in F1-optimized models.

These stage-wise analyzes reveal how predictive reliability strengthens as more information becomes available, confirming that uncertainty decreases over the course

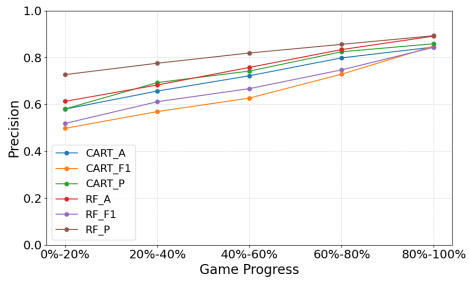
Model	Accuracy	Precision	Recall	F1 Score
CART-A	0.7648	0.7284	0.6243	0.6724
CART-F1	0.7329	0.6176	0.8114	0.7013
CART-P	0.7578	0.7520	0.5573	0.6402
RF-A	0.7760	0.7579	0.6179	0.6808
RF-F1	0.7564	0.6518	0.7943	0.7160
RF-P	0.6949	0.8383	0.2609	0.3980

(a)

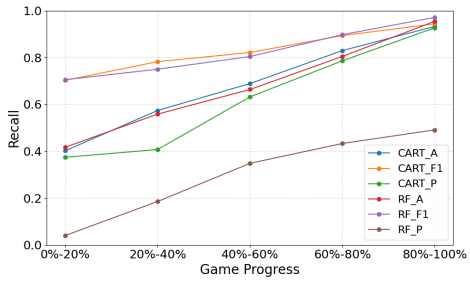


(b)

(c)



(d)



(e)

Fig. 3: Global and stage-wise evaluation across different metrics and models. (a) Overall model evaluation. (b) Accuracy across game phases. (c) F1 score across game phases. (d) Precision across game phases. (e) Recall across game phases.

of game progression. Building on these temporal dynamics, feature importances computed with scikit-learn are examined, identifying which variables most strongly drive model predictions. These global insights motivate subsequent SHAP analysis, which further characterize both global and local feature interactions.

1.4 Interpretability

Variable importance analysis revealed that the predictive power was concentrated within a small subset of features (Fig. 4). Most attributes exhibited importance values below 0.3, while a few features consistently dominated model decisions. Across all

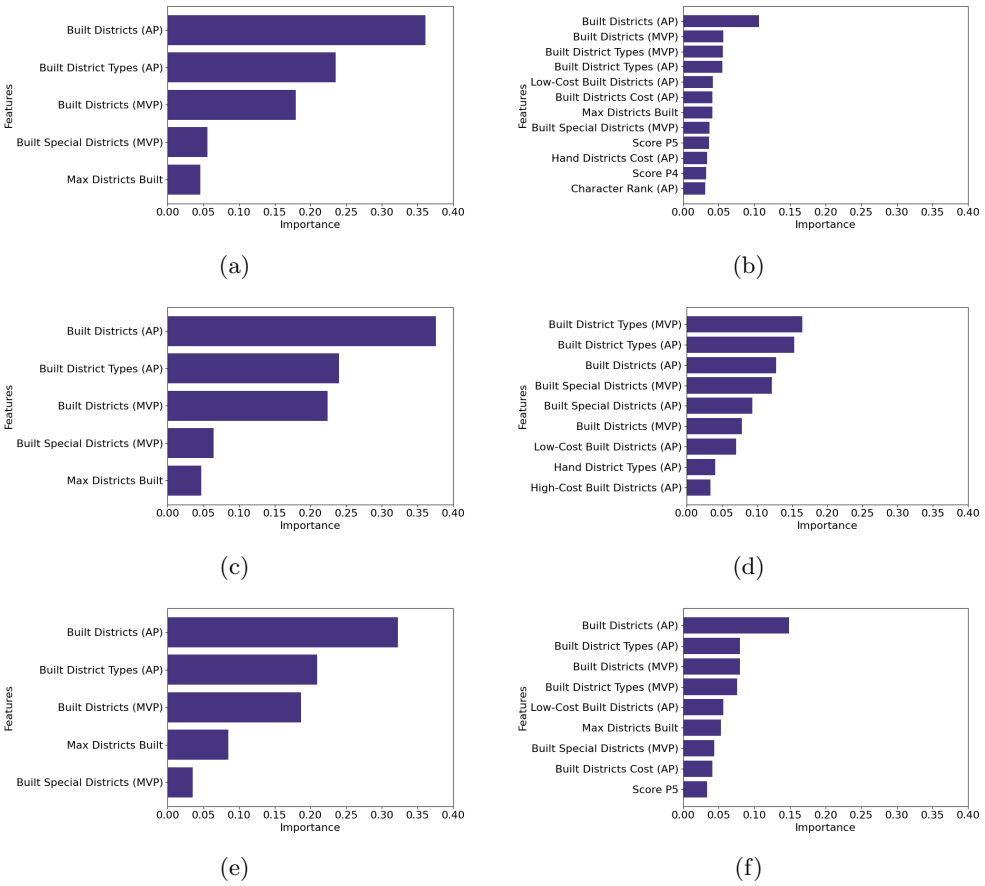


Fig. 4: Global feature importance across optimized models computed with scikit-learn CART and Random Forest (RF) classifiers optimized for (a) CART-A (b) RF-A (c) CART-P (d) RF-P (e) CART–F1 (f) RF–F1. Importance values are normalized per model to highlight the dominant predictors. Only features with importance values greater than 0.3 are shown.

models, *Built Districts (AP)* emerged as the most influential predictor, followed by *Build District Types (AP)*, indicating that the agent's ability to properly plan and execute its own constructions plays a decisive role in determining the game outcome (victory or defeat). This finding aligns with the feature–outcome correlations reported in Fig. 1c.

CART models relied on a narrower subset of variables, particularly in precision-optimized trees, which assigned weight to fewer attributes and displayed shallower structures. In contrast, accuracy-optimized Random Forests distributed importance more broadly, reflecting their ensemble flexibility and averaging effects. Precision and

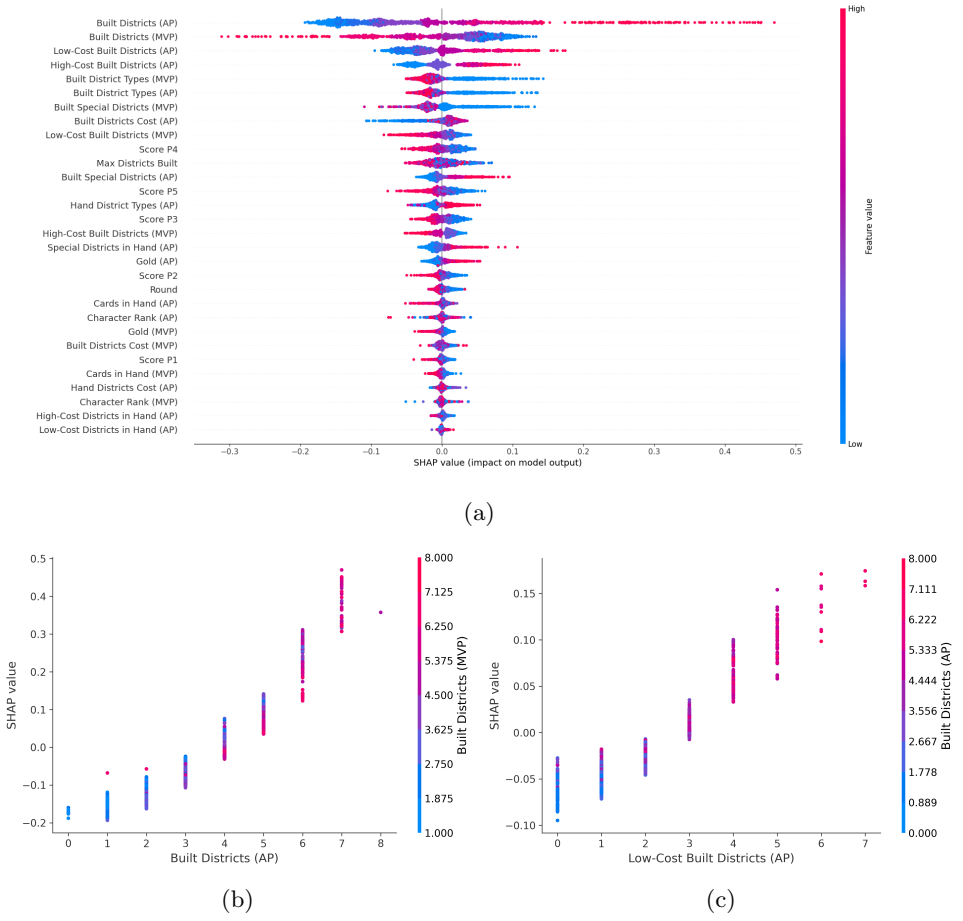


Fig. 5: SHAP analysis of feature contributions and interactions on victory predictions in the RF-A model across 1,000 randomly selected samples. (a) Beeswarm plot showing global feature effects (SHAP values) on victory predictions, with Built Districts (AP) and Built Districts (MVP) as dominant contributors. (b–c) SHAP dependence plots illustrating pairwise interactions: the influence of Built Districts (AP) is modulated by Built Districts (MVP) (b), while Low-cost Built Districts (AP) effects depend on Built Districts (AP) (c). Together, these interactions capture the relational dynamics shaping victory probabilities.

F1-optimized models emphasized features most correlated with rare victories, highlighting their ability to detect minority-class outcomes, whereas accuracy-focused models incorporated a wider set of indicators that captured general game progression.

These global importance measures are complemented by SHAP analysis, providing both global and local interpretability and uncovering feature interactions that drive individual predictions. We conducted a SHAP analysis on the most robust model,

RF-A, to examine how individual features shaped victory predictions. The beeswarm plot (Figure 5a) highlights two dominant variables: *Built Districts (AP)* and *Built Districts (MVP)*. High values of *Built Districts (AP)* strongly increased the predicted probability of victory with a SHAP value of up to 0.5, while low values reduced it to around -0.2. Conversely, *Built Districts (MVP)* primarily drove defeat, lowering victory predictions by as much as -0.3 when high and rarely exceeding 0.1 toward victory when low.

Other features also had notable effects |SHAP value| > 0.1, including *Low-cost Built Districts (AP)*, *High-cost Built Districts (AP)*, *Special Districts Built (AP)*, and *Special Districts in Hand (AP)*, which were associated with positive contributions toward victory. In contrast, features such as *Built District Types (AP and MVP)* and *Special Districts Built (MVP)* exhibited inverse contributions.

Dependence plots (Fig. 5b) reinforce this behavior. The influence of *Built Districts (AP)* increases linearly with its value but is moderated by *Built Districts (MVP)*, which counteracts predictions as both values rise. Similarly, Fig. 5c demonstrates that the effect of *Low-cost Built Districts (AP)* strengthens with higher values yet remains mediated by the central variable *Built Districts (AP)*.

Overall, the set of dominant variables remained largely consistent across both interpretability methods, with only minor differences in ranking. These results indicate that victory is primarily explained by structural city-building features, but their predictive impact depends on the relational dynamics between AP and MVP attributes. This underscores the importance of accounting for interacting strategic variables rather than isolated effects, highlighting how explainable tree-based models can reveal the balance between structure, chance, and strategy in complex multi-agent environments.

2 Discussion

Our findings demonstrate that interpretable tree-based models can accurately predict victory in a complex multi-agent environment while providing transparent insights into the mechanisms that drive success. Beyond predictive performance, these models reveal how structural, stochastic, and relational factors interact to shape strategic outcomes, linking interpretable machine intelligence to human-understandable reasoning.

Building on this foundation, the synthetic dataset and the careful selection of 30 features effectively captured the core dynamics of *Citadels*, preserving the interplay between randomness and strategic choice while minimizing noise and class imbalance (Section 1.1), while relying exclusively on publicly available features from the target player's state (AP). This representation provided a stable basis for model training and interpretability.

Leveraging this well-structured dataset, Optuna-based optimization yielded robust and reproducible trends across evaluation metrics. Precision-focused CART models converged to shallow, specialized trees, whereas accuracy-optimized Random Forests favored deeper ensembles with hundreds of estimators (Fig. 2g). These complementary optima balance computational efficiency and predictive strength. Random Forests generally outperformed CART across all metrics, though precision-oriented models

sacrificed recall, illustrating the classic trade-off between false positives and false negatives (Fig. 3). F1-optimized configurations achieved the most stable balance, particularly under early-game stochasticity. Stage-wise evaluation further revealed progressive improvement as the game advanced: accuracy and F1 score exceeded 0.85 in the late phases, and recall approached 0.9 for most models (Fig. 3b and Fig. 3c). Among all, RF-A proved consistently robust, whereas CART-P maintained steadier recall under noisy early conditions.

Feature importance and SHAP analyzes clarified the strategic drivers of victory. CART models relied on a concise set of decisive variables, while Random Forests distributed importance more diffusely across complementary signals. *Built Districts (AP)* consistently emerged as the strongest predictor, increasing victory likelihood, while *Built Districts (MVP)* reduced it. Dependence plots revealed clear interaction effects: the influence of *Built Districts (AP)* progress was moderated by *Built Districts (MVP)* districts, and the contribution of *Low-Cost Built Districts (AP)* depended on *Built Districts (AP)* progress. These relational patterns highlight that predictive accuracy arises not from isolated indicators but from the interplay of structural and competitive dynamics.

SHAP analysis provided a more context-sensitive understanding of feature relevance by capturing interdependencies and interaction effects among variables. It revealed that identical feature values can contribute differently to victory likelihood depending on the structural configuration of the state. This ability to disentangle local and global effects complements impurity-based metrics, emphasizing that predictive performance arises not from variable frequency in decision nodes, but from the dynamic interplay between structural and competitive factors shaping the outcome.

From a strategic perspective, these findings suggest that human players should prioritize consistent district construction, balance low- and high-cost districts, and treat diversity opportunistically, while leveraging role choices to either accelerate their own progress or disrupt the progress of the other players. For AI agents, SHAP profiles enable targeted evaluation of role-focused strategies to determine whether speed or economy yields more consistent outcomes. While stochasticity and interaction noise are inevitable in multi-agent games, shallow CART models capture most global behavioral patterns, whereas deeper Random Forests uncover subtle, higher-order dynamics that govern expert-level play.

Despite the synthetic nature of the dataset, the modeling framework proved robust even under randomized actions. Applying the same methodology to expert-level or tournament data would likely yield stronger and more interpretable signals, as high-skill players exploit advantages more consistently. However, real matches introduce subjectivity and behavioral volatility that synthetic simulations cannot fully reproduce. In either case, the approach provides a principled foundation for predictive modeling and explainable analysis in complex competitive domains.

By linking predictive signals to specific features, this study shows how machine intelligence can reproduce actionable insights beyond raw accuracy. Such interpretability bridges the gap between human and AI reasoning, supporting human learning, AI design, and transparent decision-making in strategic, multi-agent environments.

In summary, this work demonstrates that interpretable tree-based learning can accurately predict outcomes in complex multi-agent games while revealing the strategic mechanisms that underlie success. By linking explainable machine intelligence with structured, stochastic decision-making, we show that transparent models can serve not only as predictors but also as analytical tools for understanding competition, adaptation, and reasoning. This framework establishes a foundation for future studies combining predictive modeling, interpretability, and behavioral analysis in games and beyond—advancing the development of AI systems that are both high-performing and intrinsically explainable.

3 Methods

This section outlines the methodological framework adopted to train and evaluate tree-based learning models within the context of outcome prediction in the *Citadels* board game. It first introduces the decision tree and ensemble configurations, including hyperparameter settings and evaluation metrics. The following subsections describe the experimental environment and dataset generation, followed by the model optimization pipeline, stage-wise testing procedures, and interpretability analysis based on feature importance and SHAP values.

3.1 Decision Tree Framework

Decision trees are widely used supervised learning models, valued for their interpretability and a hierarchical structure that makes feature influence traceable [13–16]. While individual trees provide transparent rules, they are prone to overfitting, especially when they are deep or when the data is noisy. Ensemble methods, such as Random Forests, reduce variance and bias by combining multiple trees trained on bootstrap samples and random feature subsets [17]. This ensemble averaging typically improves generalization at the cost of interpretability [18].

In this study, we employed the scikit-learn [19] implementations of CART and Random Forest (RF). Model behavior is shaped by a small set of key hyperparameters:

- **max_depth**: maximum depth of each tree; deeper trees may lead to overfitting, while shallower ones may underfit.
- **criterion**: metric used to evaluate split quality; Gini impurity index, information gain (entropy), or logarithmic loss.
- **min_samples_leaf**: minimum number of samples required at a leaf node, preventing excessively specific terminal splits.
- **min_samples_split**: minimum number of samples required to split an internal node. Higher values encourage more generalized trees.
- **class_weight**: balances the relative importance of classes to mitigate the effects of class imbalance.
- **n_estimators** (RF only): number of trees in the ensemble. Larger ensembles typically improve performance at the cost of a higher computational cost and reduced interpretability.

Hyperparameters directly influence the bias–variance trade-off. Deep trees or large ensembles tend to capture more complex interactions but risk overfitting, while shallow trees promote generalization at the expense of expressiveness. Effective tuning, therefore, requires systematic optimization guided by multiple performance metrics.

3.1.1 Evaluation Metrics

Because the dataset is moderately imbalanced, relying solely on accuracy may obscure meaningful differences between models. Complementary metrics such as precision, recall, and F1 score provide a fuller view of performance under asymmetric class distributions. Precision captures false-positive control, recall quantifies sensitivity to rare victories, and F1 score balances both [20–22].

All evaluations employed stratified cross-validation to preserve class proportions and produce reliable estimates of generalization. This combination of robust metrics and stratified sampling ensures a consistent comparison between optimization objectives.

This foundation motivates the choice of *Citadels* as the experimental testbed. Its multi-agent, stochastic, and strategically rich environment allows for the assessment of interpretable models, such as decision trees, evaluating both predictive performance and structural interpretability in a complex scenario.

3.2 The *Citadels* Environment

Citadels (Designed by Bruno Faidutti in 2000; revised edition launched in 2016 [23]), also known as *Machiavelli* in Europe, is a multiplayer (2-8 players), turn-based strategy board game that combines economic planning and hidden roles through closed drafting and set-collection mechanics, set within a medieval city-building theme, as illustrated in Supplementary Fig. 6a and 6b.

Players build cities by constructing districts of varying costs and types while using character abilities to disrupt opponents or accelerate their own progress. The game ends when one player completes seven districts, and victory is determined by total city value, diversity bonuses, and special district effects [23].

This dynamic, partially observable, and stochastic environment offers a rich testbed for evaluating machine learning models in complex decision spaces that require long-term planning. The interplay of deterministic rules and probabilistic events closely reflects real-world decision-making under uncertainty, making *Citadels* a suitable domain for studying interpretable, performance-oriented predictive systems.

3.3 Dataset Generation

Due to the limited availability of structured human gameplay data, a synthetic dataset was generated using rule-based agents. Synthetic approaches are widely recognized for enabling reproducibility and controllable complexity [24,25], though they must be carefully designed to capture realistic variability [26,27].

Each simulated match involved five players, which is the most recommended configuration according to the BoardGameGeek[1] community. Six rule-based agents

[1] https://boardgamegeek.com/boardgame/478/citadels

represented distinct strategic human-like archetypes, complemented by one fully stochastic Random agent to introduce behavioral noise. All possible agent combinations (with repetition) were played in a round-robin tournament of 25 matches each, producing a large and diverse dataset of partial game states labeled by the active player's (AP) eventual outcome.

Thirty features were selected to represent both structural and strategic aspects of gameplay. Each state representation was encoded from the perspective of the active agent (AP), incorporating its personal data and the information known by the AP about the player with the highest partial score (MVP). When the AP coincided with the MVP of the match, the second-highest-scoring player was used instead. Observing the MVP reflects a common strategy among human players in competitive multiplayer games, as monitoring the leading opponent often informs adaptive decision-making. Partial scores reflected the cumulative cost of districts up to that round, thereby capturing evolving strategies rather than final outcomes alone. This abstraction preserves diverse interactions while emphasizing dynamics.

This dataset served as the basis for hyperparameter optimization using Optuna [28], tuning decision-tree–based models across multiple evaluation metrics, as described in the following subsection.

3.4 Hyperparameter Optimization

Hyperparameters were optimized using Optuna [28], a Bayesian optimization framework that efficiently explores continuous search spaces. Each optimization study included up to 1,000 trials with pruning and early stopping (patience = 100) to minimize redundant computation [29,30]. Broader parameter ranges were defined around the defaults of the scikit-learn [19] (Supplementary Table 1), ensuring the exploration of high-capacity and lightweight model configurations.

Each study targeted a specific metric (accuracy, precision, or F1 score), while recall was used exclusively for model evaluation. Five-fold stratified cross-validation ensured balanced data splits. Class weights were adjusted to emphasize the minority class (victory) and mitigate the effects of imbalanced classes.

For each optimized model, we recorded the convergence patterns and alternative near-optimal configurations. The resulting models were subsequently tested under consistent conditions to evaluate their generalization and robustness.

Building on these results, we performed a comprehensive evaluation encompassing overall performance, optimized hyperparameters, scikit-learn feature importances, and stage-wise assessments across game phases. This analysis provides a holistic view linking structural configurations to predictive and interpretative performance.

3.5 Stage-wise Evaluation

Initially, models were evaluated globally using overall performance metrics (accuracy, precision, recall, and F1 score) to provide a general assessment of predictive capability. However, such aggregate results can mask temporal variability across the game.

To evaluate robustness over time, models were re-assessed using new synthetic samples distributed across five relative intervals in relation to the game progression:

$[0\%, 20\%)$, $[20\%, 40\%)$, $[40\%, 60\%)$, $[60\%, 80\%)$, and $[80\%, 100\%]$. Each test set contained 10 combinations per match configuration and followed the same guidelines described for the original dataset.

Performance metrics were computed at each interval to quantify how predictive reliability evolved as the game advanced. This stage-wise analysis links predictive strength to game dynamics, establishing a foundation for interpretability assessment.

3.6 Interpretability Analysis

Interpretability is central to understanding and validating predictive models in complex, multi-agent domains. While decision trees are inherently transparent, ensembles like Random Forests obscure individual decision paths, necessitating model-agnostic approaches to explain their behavior.

Feature importances were computed using the native `scikit-learn` implementation. This approach estimates variable relevance through the mean decrease in impurity (MDI), which quantifies the average reduction in node impurity measured by the chosen split criterion and weighted by the number of samples reaching each node. Consequently, features that consistently contribute to more informative splits receive higher importance scores. Although inherently model-specific, this metric provides an efficient and interpretable approximation of each feature's global influence and enables cross-comparison between CART and Random Forest models.

Additionally, we employed SHAP (SHapley Additive exPlanations) [6] with the TreeExplainer framework [31], which decomposes each prediction into additive feature contributions. SHAP values capture both global importance and local variability, revealing feature interactions that are particularly relevant in *Citadels*, where structural progress (district construction) interacts with dynamic strategic factors such as character selection and resource management.

Together, these analyzes link quantitative performance with strategic meaning, clarifying not only why the models predicted victories but also how their reasoning aligns with observable game dynamics.

Supplementary information.

Acknowledgements. The authors used OpenAI's ChatGPT (GPT-5 model, 2025) to assist in revising the language and improving the clarity of the manuscript. The authors reviewed and edited the content to ensure accuracy and accountability for all statements.

Data and Code Availability. All data, trained models, and source code supporting the findings of this study are publicly available at: anonymized. The repository includes the controlled game environment, agent implementations, and all scripts required for model training and evaluation. Specifically, the branch `anonymized` contains the code and analyzes used to reproduce the experiments reported in this work.

References

[1] Fürnkranz, J.: Machine learning in games: A survey. In: Machine Learning in Games: A Survey, pp. 11–59 (2001)

[2] Breiman, L.: Classification and Regression Trees. Routledge, London, UK (2017)

[3] Pandolfini, B.: Kasparov and Deep Blue: The Historic Chess Match Between Man and Machine. Simon and Schuster, New York, USA (1997)

[4] Silver, D., *et al.*: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016) https://doi.org/10.1038/nature16961

[5] Silver, D., *et al.*: A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science **362**(6419), 1140–1144 (2018) https://doi.org/10.1126/science.aar6404

[6] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[7] Ge, Z., *et al.*: Concept learning for cooperative multi-agent reinforcement learning. In: 2025 IEEE 26th China Conference on System Simulation Technology and Its Applications (CCSSTA), pp. 625–631 (2025). https://doi.org/10.1109/IEEECONF65522.2025.11136965

[8] Billings, D., *et al.*: Using probabilistic knowledge and simulation to play poker. In: AAAI (1999)

[9] Dockhorn, A., *et al.*: Predicting opponent moves for improving hearthstone ai. In: IPMU (2018)

[10] OpenAI: OpenAI Five Defeats Dota 2 World Champions (2019). https://openai.com/index/openai-five-defeats-dota-2-world-champions/

[11] Do, T.D., et al.: Using Machine Learning to Predict Game Outcomes Based on Player-Champion Experience in League of Legends. 16th International Conference on Foundations of Digital Games (2021)

[12] Junior, J.B.S., Campelo, C.E.C.: League of legends: Real-time result prediction. In: Simas, E., Ferreira, D.D., Oliveira, L.R. (eds.) Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC'2023), pp. 1–8. SBIC, Salvador, BA (2023). https://doi.org/10.21528/CBIC2023-161

[13] Blockeel, H., et al.: Decision trees: from efficient prediction to responsible ai. Sec. Machine Learning and Artificial Intelligence **6** (2023) https://doi.org/10.3389/frai.2023.1124553

[14] Ville, B.: Decision trees. WIREs Computational Statistics **5**(6), 448–455 (2013)

https://doi.org/10.1002/wics.1278

[15] Rokach, L., Maimon, O.: In: Maimon, O., Rokach, L. (eds.) Decision Trees, pp. 165–192. Springer, Boston, MA (2005). https://doi.org/10.1007/0-387-25465-X_9

[16] Jiang, T., *et al.*: Supervised machine learning: A brief primer. Behavior Therapy **51**(5), 675–687 (2020) https://doi.org/10.1016/j.beth.2020.05.002

[17] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[18] Eryarsoy, E., Delen, D.: Predicting the outcome of a football game: A comparative analisys of single and ensemble aanalytics methods. Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)

[19] Pedregosa, F., *et al.*: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[20] Hicks, S.A., *et al.*: On evaluation metrics for medical applications of artificial intelligence. Scientific Reports **12**(1), 5979 (2022) https://doi.org/10.1038/s41598-022-09954-8

[21] Hand, D.J., *et al.*: F*: an interpretable transformation of the f-measure. Machine Learning **110**(3), 451–456 (2021) https://doi.org/10.1007/s10994-021-05964-1

[22] Narwane, S.V., Sawarkar, S.D.: Machine learning and class imbalance: A literature survey. Ind. Eng. J **12**(10.26488) (2019)

[23] Faidutti, B.: Citadels, Deluxe ed. edn. Z-MAN games, Roseville, MN, USA (2016). Z-MAN games. Rulebook. https://images-cdn.zmangames.com/us-east-1/filer_public/82/aa/82aac2d6-2a19-4143-9690-eb16b82bd9af/citadels_deluxe_rulebook.pdf

[24] Nam, G., *et al.*: Gcisg: Guided causal invariant learning for improved syn-to-real generalization. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, pp. 656–672. Springer, Cham (2022)

[25] Nguyen, L.-C., et al.: Provably Improving Generalization of Few-Shot Models with Synthetic Data (2025). https://arxiv.org/abs/2505.24190

[26] Ba, Y., *et al.*: Fill in the gaps: Model calibration and generalization with synthetic data. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 17211–17225. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.955

[27] Hosseini, M.J., *et al.*: A synthetic data approach for domain generalization of

NLI models. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2212–2226. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.120

[28] Akiba, T., *et al.*: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631 (2019)

[29] Lai, J.-P., *et al.*: Tree-based machine learning models with optuna in predicting impedance values for circuit analysis. Micromachines **14**(2), 265 (2023)

[30] Duță, S., Sultana, A.E.: Optimizing depression classification using combined datasets and hyperparameter tuning with optuna. Sensors **25**(7), 2083 (2025)

[31] Lundberg, S.M., *et al.*: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**(1), 2522–5839 (2020)

[32] Faidutti, B.: Citadels. Board game. Multisim, France. BoardGameGeek entry: https://boardgamegeek.com/boardgame/205398/citadels (2000). https://boardgamegeek.com/boardgame/205398/citadels

Hyperparameter	Search space
max_depth	[2, 50], None
min_samples_leaf	[2, 500]
min_samples_split	[2, 500]
class_weight	0:1, 1:5,4,3,2,1
criterion	['gini', 'entropy', 'log_loss']
n_estimators*	[50, 1000]

** Applicable only to Random Forest.*

Table 1: Supplementary Table 1: Hyperparameter search space for CART and Random Forest models.

(a) (b)

Fig. 6: **Supplementary Figure 1** — Elements of the Citadels game. (a) Box from the 2016 revised edition. (b) Components including character cards, district cards and gold tokens. Images reproduced from [32].