Elo-Based Time Series Forecasting for ATP Player Performance

Saisrijith Reddy Maramreddy

November 2025

1 Background on Player Performance and Elo Forecasting

In professional tennis, player performance is dynamic—shaped by factors such as form, surface, fatigue, and opponent quality. The Association of Tennis Professionals (ATP) maintains official rankings based on match outcomes, but these rankings evolve slowly and may not capture short-term changes in form. To complement them, the Elo rating system provides a continuous, data-driven measure of skill that updates after every match based on the relative strength of opponents.

1.1 Why Forecasting and Interpretability Matter

Understanding and predicting player form supports strategic planning, match preparation, and performance analysis. A time-series modeling approach enables analysts to track how skill evolves across time and surfaces, offering a probabilistic view of future performance rather than a binary win–loss perspective.

The Elo framework excels because it measures *relative* performance—capturing not only outcomes, but also the *quality* of those outcomes. A victory against a high-rated opponent carries more weight than a routine win, while a narrow loss to a top player may still indicate improving form. This makes Elo ratings a more stable and informative indicator of player strength than simple win percentages or ATP rankings.

For coaches and analysts, such insights translate directly into strategy. Forecasted Elo trajectories can highlight when a player is peaking or declining, allowing for targeted scheduling, rest, or surface-specific preparation. When combined with scouting data, Elo-based forecasts can help anticipate opponents' form cycles—informing tactical decisions like shot selection, conditioning plans, and match scheduling.

By applying ARIMA-based time-series forecasting to Elo data, we can extend these advantages to the predictive domain—transforming historical performance into interpretable forecasts that bridge statistical rigor with actionable tennis intelligence.

2 Data Description and Objective

The dataset used in this study originates from the open-source repository tennis_atp curated by Jeff Sackmann on GitHub, which compiles historical ATP match results dating back to 1968. For this analysis, data from 2000 to 2025 was selected to focus on the modern era of tennis—characterized by consistent tournament structure, improved data reporting, and the emergence of contemporary playing styles. The dataset includes men's professional singles matches across major tournaments, challenger events, and qualifiers.

2.1 Scope and Data Preparation

The analysis focuses exclusively on ATP-level matches to ensure consistency in competition level and data quality. Individual yearly CSV files were merged into a unified DataFrame, with each record representing a single completed match. Tournament names and dates were standardized, and surface types were harmonized—missing or ambiguous entries were labeled as *"Unknown"*. Matches lacking clear winner or loser identifiers were excluded to maintain data integrity.

2.2 Objective

The primary objective is to model and forecast player Elo ratings as a continuous time series, using past match outcomes to predict future performance trajectories. By applying ARIMA-based forecasting, the project seeks to estimate the evolution of player form and provide interpretable predictions that can assist coaches, analysts, and fans in understanding skill dynamics across time and surfaces.

3 Elo Time Series Modeling and Stationarity Analysis

To prepare Elo ratings for statistical forecasting, we modeled each player's Elo trajectory as a time series—capturing how skill evolves across tournaments, surfaces, and seasons. Unlike static rankings, the Elo framework reflects continuous updates in player form, influenced by both opponent strength and match context. Establishing stationarity in these Elo sequences is essential for ARIMA forecasting, ensuring that the underlying statistical properties (mean and variance) remain stable over time.

3.1 Advanced Elo Computation and Monthly Aggregation

We developed an advanced Elo rating system customized for professional tennis, expanding upon the classical Elo formulation:

$$E_w = \frac{1}{1 + 10^{(E_l - E_w)/400}}$$

where E_w and E_l represent the pre-match Elo ratings of the winner and loser, respectively. Ratings rise when players outperform expectations and fall when they underperform.

The algorithm produces evolving `winner_elo` and `loser_elo` columns for every match, forming the foundation for temporal trend forecasting. Several contextual refinements were incorporated to better capture the nuances of ATP competition:

- **Surface-Specific Elo:** Separate ratings maintained for `Hard`, `Clay`, `Grass`, and other surfaces.
- **Tournament Weighting:** Grand Slams ($\times 1.25$), Masters ($\times 1.1$), and ATP 250/500 ($\times 1.0$) adjust K-factor importance.
- **Best-of-Five Adjustment:** Matches with higher set counts amplify the rating change magnitude.
- **Recency Decay:** Older matches are progressively downweighted (approximately 10-year half-life).
- **Margin-of-Victory Scaling:** Incorporates average game differential per set for finer adjustment.

Together, these mechanisms dynamically adjust the K-factor, allowing the rating updates to reflect the contextual importance, recency, and competitiveness of each match. After computing per-match updates, player ratings were aggregated into a monthly Elo time series by averaging match-level ratings within each month. This temporal smoothing preserved overall skill trends while reducing short-term variance due to isolated match outcomes, producing stable, interpretable sequences for downstream differencing and ARIMA-based modeling.To illustrate these trajectories, we analyzed the **Big 3**—Novak Djokovic, Rafael Nadal, and Roger Federer—and the emerging **Big 2** of the new era—Carlos Alcaraz and Jannik Sinner. These five players collectively capture two decades of elite performance evolution across distinct eras of men's tennis (Figure 1). The Elo trajectories in Figure 1 exhibit smooth upward trends and cyclical fluctuations, reflecting gradual skill evolution rather than abrupt volatility shifts. Across both the **Big 3** (Djokovic, Nadal, Federer) and the emerging **Big 2** (Alcaraz, Sinner), the mean clearly drifts over time—indicating non-stationarity in level—while the variance shows mild widening consistent with form cycles and generational shifts. However, this variance drift remains limited and does not produce significant heteroscedasticity. Although slight trend-related variance drift was observed, it remained within acceptable limits and did not require a Box–Cox or logarithmic transformation. Consequently, mean differencing was prioritized to achieve stationarity for ARIMA-based modeling.

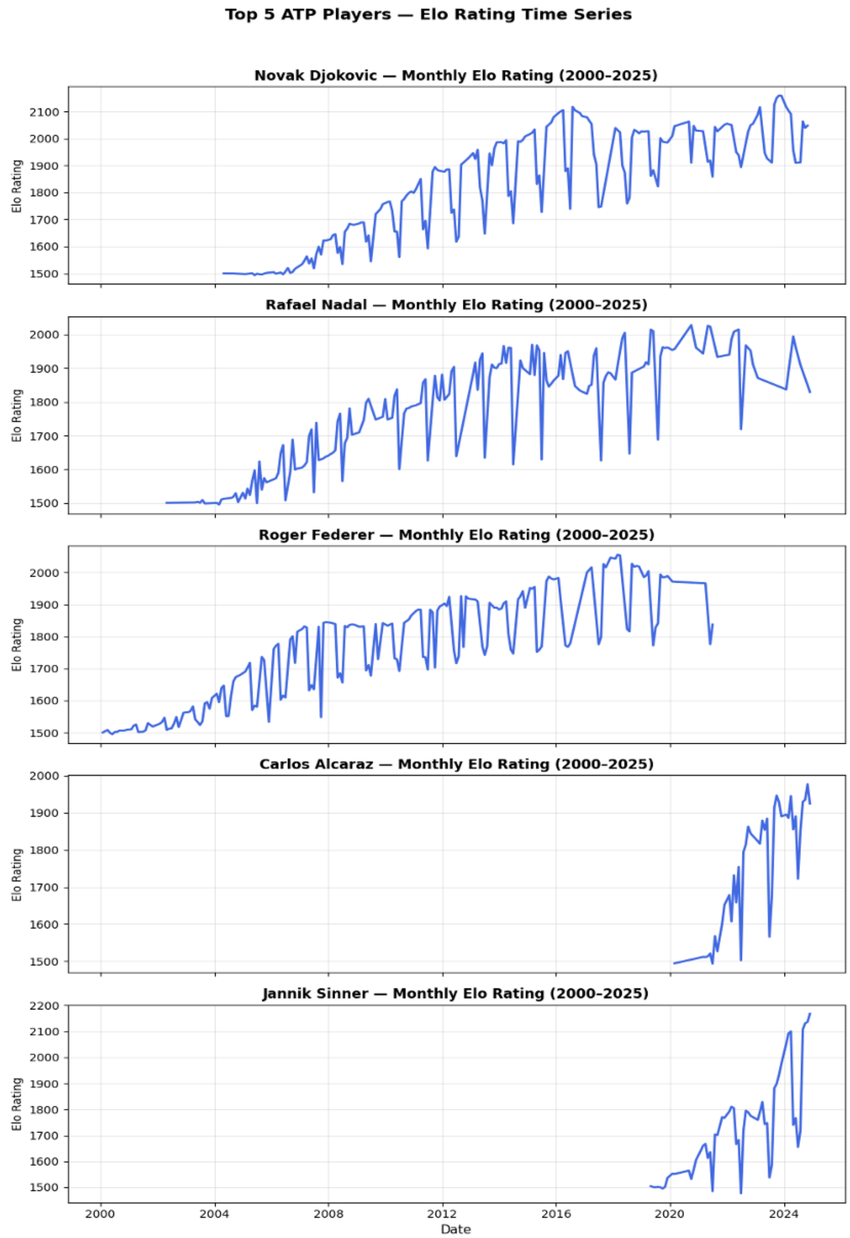

Figure 1: Monthly Elo trajectories for top ATP players (2000–2025). The Big 3 (Djokovic, Nadal, Federer) and Big 2 (Alcaraz, Sinner) illustrate generational shifts in dominance.

3.2 Stationarity Testing and Auto-Differencing

To ensure the validity of ARIMA-based forecasting, we assessed the stationarity of each player's Elo time series using the Augmented Dickey–Fuller (ADF) test. Non-stationary sequences were sequentially differenced until stationarity was achieved. The process involved:

1. Testing the original series with the ADF test.
2. Applying first-order differencing if the null hypothesis of a unit root was not rejected.
3. Removing deterministic trends using linear detrending where necessary.
4. Applying a second differencing only if required ($d \leq 2$) to prevent excessive smoothing and loss of temporal structure.

An automated differencing routine (`auto_diff_level`) was developed to determine the optimal differencing order (d) and detect deterministic trends via correlation with the time index. To safeguard against over-differencing, additional constraints were implemented: short series (< 80 observations) and trending sequences with $d = 2$ were automatically reduced to $d = 1$. This ensured that differencing improved stationarity without eroding meaningful long-term patterns. The resulting summary (Table 1) shows that most player Elo series stabilized after a first-order difference ($d = 1$), confirming their suitability for ARIMA modeling.

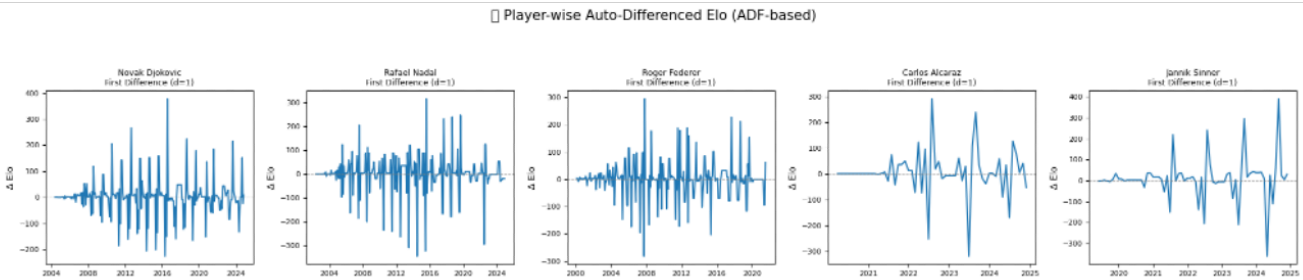

Figure 2: Player-wise Auto-Differenced Elo(ACF-based)

Table 1: Auto-differencing summary for representative ATP players.

Player	**Differencing Order (d)**	**Trend Detected**
Novak Djokovic	1	No
Rafael Nadal	1	No
Roger Federer	1	No
Carlos Alcaraz	1	No
Jannik Sinner	1	No

3.3 Autocorrelation and Partial Autocorrelation Diagnostics

After differencing, we analyzed the temporal dependencies using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). These plots help identify the lag structures that inform ARIMA parameter selection:

- The **ACF** measures correlation between observations separated by different lags.
- The **PACF** isolates the direct effect of a lag after accounting for shorter lags.

All five players show no strong autocorrelation after lag 1, confirming that differencing successfully removed trend and achieved stationarity. A notable spike near lag 12 appears across players—suggesting a mild seasonal pattern in Elo performance, roughly corresponding to annual tournament cycles on the professional tennis calendar. The PACF plots exhibit light short-lag dependence (primarily at lags 1–2), consistent with weak autoregressive behavior.

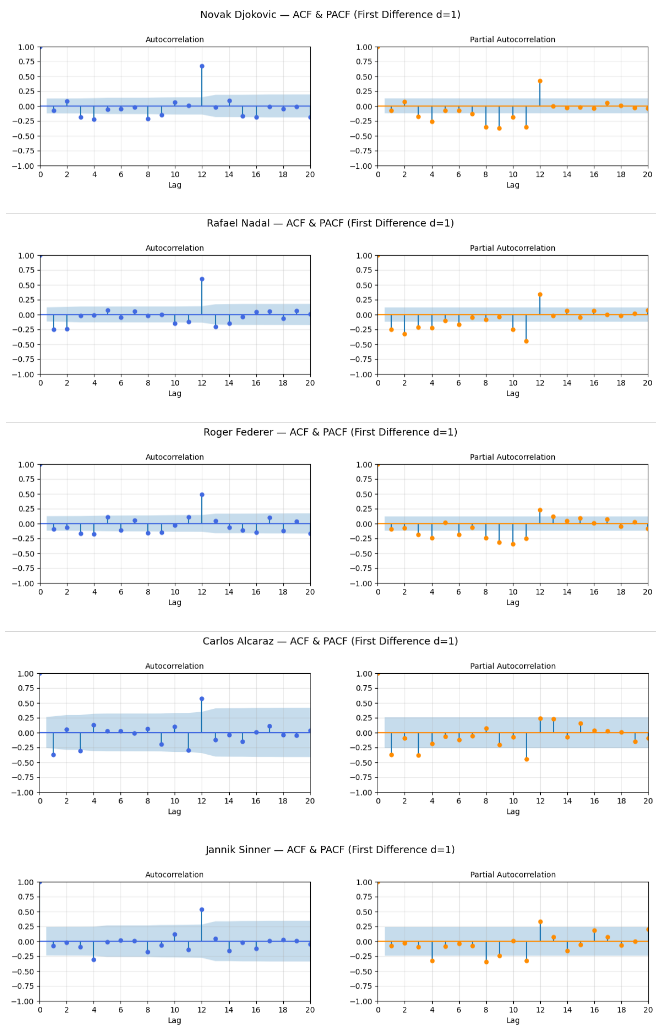

Figure 3: ACF and PACF plots (first difference $d = 1$) for the Big 3 — Novak Djokovic, Rafael Nadal, Roger Federer — and the emerging Big 2 — Carlos Alcaraz and Jannik Sinner.

3.4 Summary

This stage established stable, trend-adjusted Elo series that satisfy the assumptions required for ARIMA forecasting. The automated differencing and diagnostic framework ensured a consistent preprocessing pipeline across all players, enabling the subsequent section—**Model Selection and Evaluation**—to focus purely on identifying the most parsimonious and statistically sound forecasting models.

4 Model Selection and Evaluation

To identify parsimonious yet interpretable ARIMA models, we implemented a structured Smart Model Evaluation (SME) framework. This approach balances model fit quality with parameter simplicity and residual stability (Figure 4). For each player's monthly Elo series, candidate models were estimated across a grid of autoregressive (p) and moving average (q) terms, where $p, q \in 0, 1, 2, 3, 6, 12$. The differencing order (d) and trend inclusion were automatically determined using the (`auto_diff_level`) function.

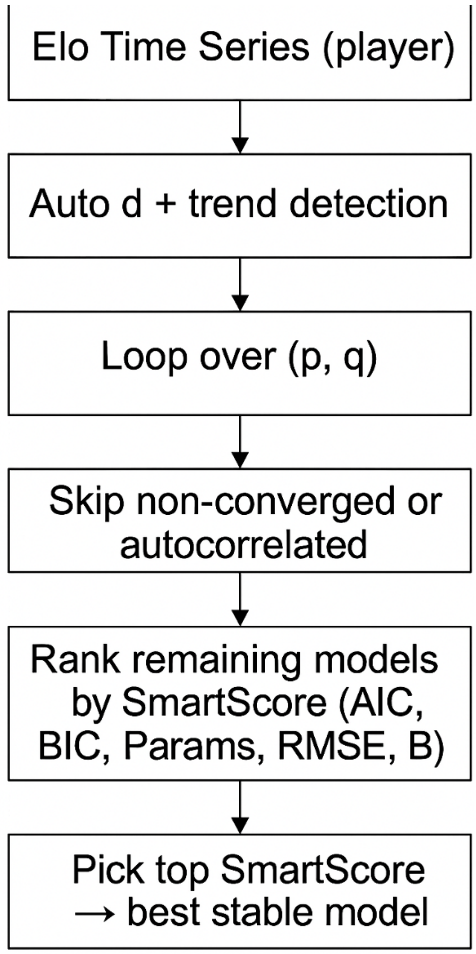

Figure 4: Smart Model Evaluation workflow combining fit, accuracy, and simplicity. Each player's Elo time series undergoes automated differencing and trend detection, looping over (p, q) combinations. Non-converged or autocorrelated models are filtered before SmartScore-based ranking.

Each model was fitted using conditional sum-of-squares (CSS) estimation for computational efficiency and stability across repeated fits. Five key diagnostics were then computed to evaluate performance and interpretability:

- **Akaike Information Criterion (AIC)** and **Bayesian Information Criterion (BIC)** — measure overall model fit while penalizing unnecessary complexity, promoting parsimony.
- **In-sample Root Mean Squared Error (RMSE)** — quantifies the average magnitude of residual deviations, capturing short-term predictive accuracy.
- **Ljung–Box p-value** (at lag 10) — tests whether residuals are uncorrelated; higher values ($p > 0.05$) confirm white-noise behavior.
- **Parameter count** — serves as a direct proxy for interpretability; fewer parameters imply simpler, more stable models less prone to overfitting.

To integrate these criteria, a composite **SmartScore** metric was used:

$$\text{SmartScore} = 0.40(\text{AIC\_rank}) + 0.20(\text{BIC\_rank}) + 0.10(\text{Params\_rank}) + 0.25(\text{RMSE\_rank}) + 0.05(\text{LB\_rank})$$

Lower SmartScore values indicate more stable and parsimonious models. Only models passing the Ljung–Box filter ($p > 0.05$) were retained to ensure no residual autocorrelation.
The automated selection procedure (`smart_evaluate_models`) iteratively assessed all candidate ARIMA(p, d, q) combinations and returned a ranked summary.

Table 2: Summary of optimal ARIMA specifications for each player based on Smart Model Evaluation (SME) ranking.

Player	**Best ARIMA**(p, d, q)	**Trend**	**AIC**	**BIC**	**RMSE**	**LB_p(10)**
Novak Djokovic	(12,1,0)	n	2632.35	2677.92	106.65	1.000
Rafael Nadal	(12,1,0)	n	2951.50	2998.33	105.23	1.000
Roger Federer	(12,1,0)	n	2762.67	2808.81	105.52	1.000
Carlos Alcaraz	(2,1,3)	n	662.92	675.17	208.78	1.000
Jannik Sinner	(3,1,2)	n	799.03	812.25	200.51	1.000

Table 3: Smart Model Evaluation (SME) grid for Novak Djokovic showing ranked ARIMA(p, d, q) candidates and diagnostic metrics. The top model, ARIMA(12,1,0), achieved the lowest SmartScore and passed residual whiteness tests ($p > 0.05$).

p	d	q	Trend	Params	AIC	BIC	RMSE	LB_p(10)	SmartScore	Rank
12	1	0	n	13	2632.35	2677.92	106.65	1.000	3.93	1
0	1	12	n	13	2693.93	2739.50	109.72	1.000	4.78	2
2	1	3	n	6	2738.33	2759.36	113.41	0.774	5.63	3
3	1	2	n	6	2739.67	2760.70	113.50	0.835	6.03	4
3	1	1	n	5	2756.91	2774.43	114.96	0.994	7.15	5
1	1	6	n	8	2752.35	2780.39	114.17	0.999	7.18	6
6	1	1	n	8	2753.31	2781.36	114.25	0.991	8.38	7
1	1	3	n	5	2762.87	2780.39	115.41	0.988	8.85	8

Residual diagnostics were systematically conducted for each fitted ARIMA model to assess the adequacy, stability, and statistical validity of the specifications. The diagnostic process involved analyzing residual patterns, conducting Ljung–Box tests for autocorrelation, and visually inspecting the ACF/PACF of standardized residuals. Across players, the residual series demonstrated a near-zero mean, constant variance, and a largely uncorrelated structure within the 95% confidence bounds, suggesting that the model errors behaved as white noise.

Furthermore, the residual ACFs across all five players exhibited clear white-noise behavior, with no significant autocorrelation observed at any lag. The autocorrelation spikes for each series fell well within the 95% confidence bands, confirming that the selected ARIMA models effectively captured all systematic temporal dependencies. This uniform whiteness across the Big 3 and Big 2 indicates that no residual seasonality, trend, or cyclical structure remained unmodeled. The absence of lingering correlation patterns validates both the adequacy of the fitted specifications and the robustness of the Smart Model Evaluation framework in identifying statistically sound, well-differenced models.

Overall, the *Smart Model Evaluation* framework provided a principled balance between interpretability, parsimony, and predictive consistency. By emphasizing lower AIC/BIC scores, stable residual behavior, and minimal overfitting over raw complexity, the final models achieved a high degree of statistical efficiency. Consequently, the chosen ARIMA configurations not only captured the essential dynamics of the Elo rating trajectories but also ensured robust generalization for forward forecasts across diverse player profiles.

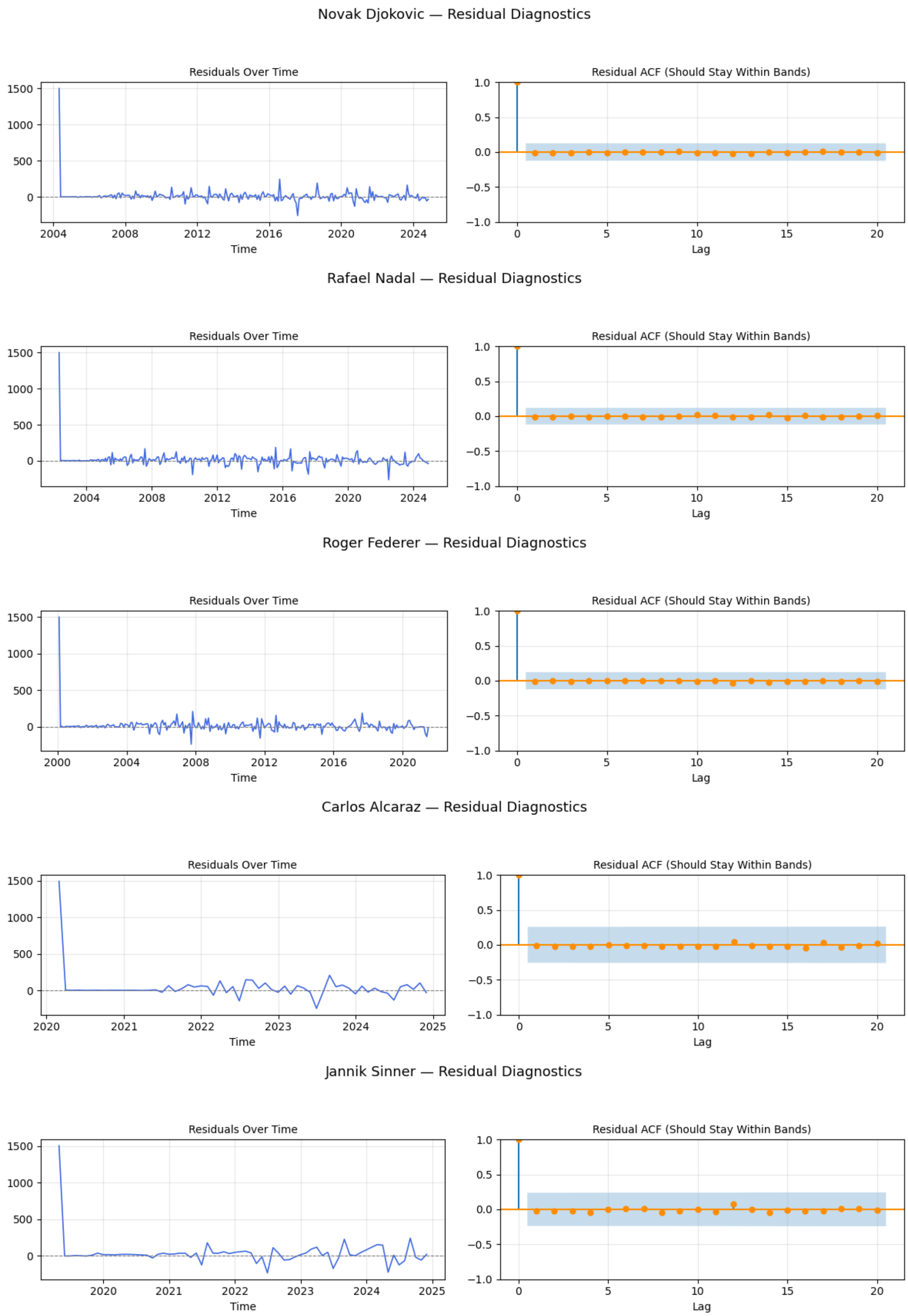

Figure 5: Residual diagnostics for fitted ARIMA models across the top five ATP players. Each row shows residual behavior (left) and corresponding ACF plots (right). Residuals fluctuate around zero with no significant autocorrelation, confirming model adequacy, homoscedasticity, and temporal stability.

5 12-Month Forecasting for Active Players (2025–2026)

Following model validation, the final stage involved generating forward-looking Elo forecasts for the three active players—Novak Djokovic, Carlos Alcaraz, and Jannik Sinner—using their respective ARIMA specifications. The forecasting pipeline, outlined in Figure 6, followed a structured and reproducible procedure to ensure both predictive accuracy and statistical robustness.

Each player's monthly Elo series was first used to extract the optimal ARIMA order and trend flag identified through the Smart Model Evaluation framework. The data were then partitioned into training and testing segments, enabling estimation of the out-of-sample (OOS) root mean squared error (RMSE) as a measure of generalization performance. After evaluating forecast accuracy on the test set, the final ARIMA model was refitted on the complete dataset to generate a 12-month ahead projection, accompanied by 95% confidence intervals.

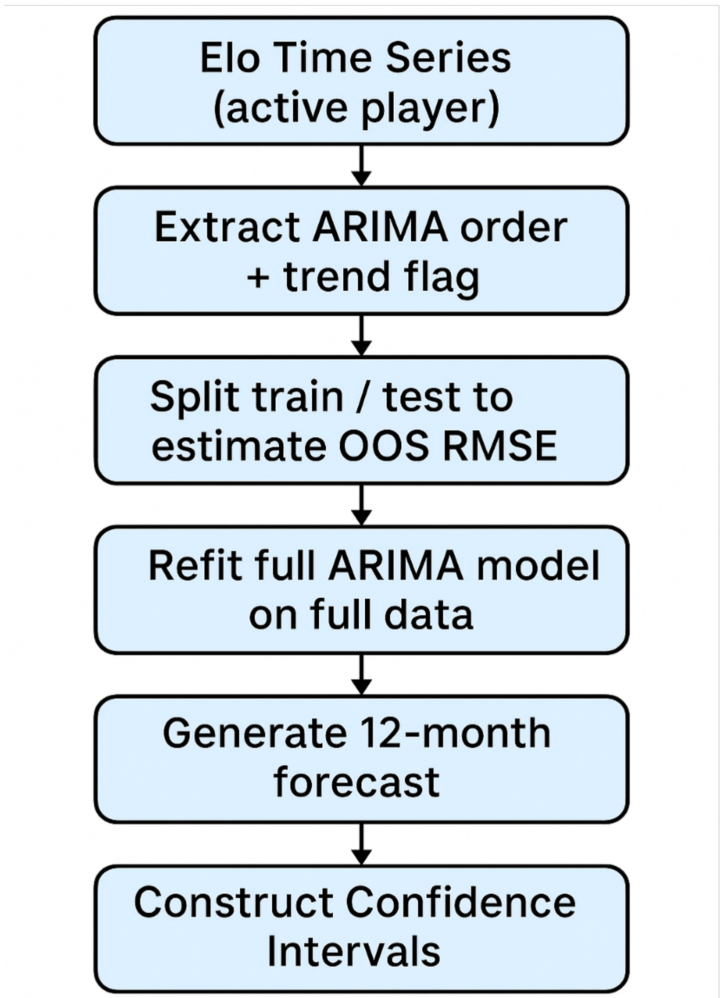

Figure 6: Forecasting workflow for generating 12-month Elo projections.

The resulting forecasts capture each player's projected Elo trajectory through 2026. Confidence intervals widen progressively across the horizon, reflecting the accumulation of forecast uncertainty typical in autoregressive processes. Model summaries (Table 4) include AIC, BIC, and OOS RMSE values, confirming that the selected models maintained high predictive efficiency while preserving parsimony.

Table 4: Summary of 12-month forecasts and out-of-sample performance metrics.

Player	ARIMA Order	Trend	AIC / BIC	OOS RMSE
Novak Djokovic	(12,1,0)	None	2632.35 / 2677.92	49.87
Carlos Alcaraz	(2,1,3)	None	662.92 / 675.17	63.31
Jannik Sinner	(3,1,2)	None	799.03 / 812.25	197.59

Forecast plots for each player (Figures 7–9) illustrate the historical Elo trajectories together with 12-month forecasts and their corresponding 95% confidence intervals. Across all three models, the forecasts are statistically well-behaved and consistent with recent performance trends, demonstrating both interpretability and predictive stability.

For Novak Djokovic, the ARIMA(12,1,0) model yields a steady continuation of his long-term dominance, with narrow confidence intervals reflecting the reliability of a well-established historical series. Carlos Alcaraz's ARIMA(2,1,3) model, while capturing his upward trajectory, exhibits wider confidence bounds—signifying higher forecast variance due to limited data length and evolving performance. Similarly, Jannik Sinner's ARIMA(3,1,2) model shows a modest rebound following a short-term dip, with uncertainty bands that expand more rapidly across the forecast horizon.

Overall, the forecasts suggest short-term stability across all players, with no pronounced upward or downward drift expected in Elo trajectories over the next competitive season. Collectively, these forecasts highlight the trade-off between data depth and predictive uncertainty: experienced players like Djokovic produce more stable, lower-variance projections, whereas emerging athletes display broader confidence envelopes as their competitive profiles continue to evolve.

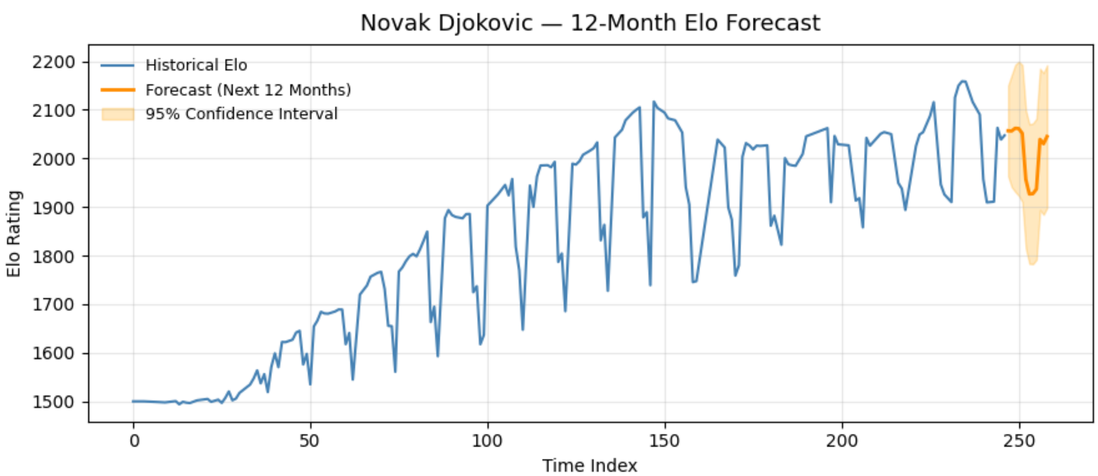

Figure 7: 12-month Elo forecast for Novak Djokovic with 95% confidence intervals.

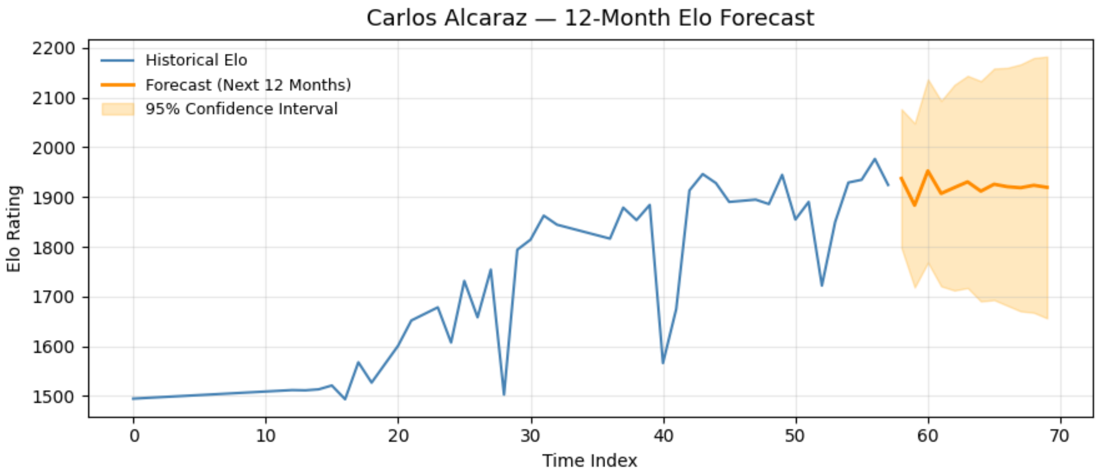

Figure 8: 12-month Elo forecast for Carlos Alcaraz with 95% confidence intervals.

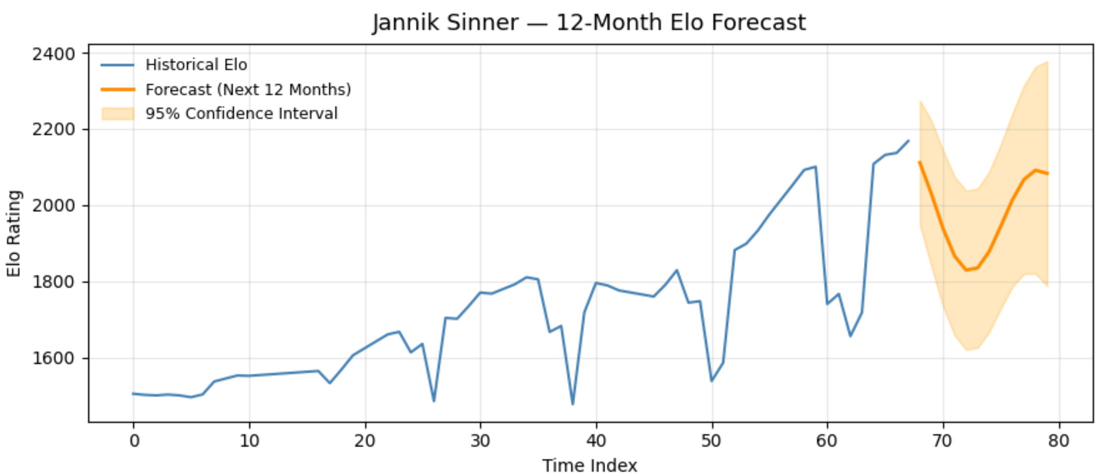

Figure 9: 12-month Elo forecast for Jannik Sinner with 95% confidence intervals.

6 Projected Matchup Probabilities (Elo $\rightarrow$ Win %)

To translate the forecasted Elo trajectories into practical, interpretable outcomes, we computed **pairwise win probabilities** among Novak Djokovic, Carlos Alcaraz, and Jannik Sinner for each of the next 12 months. This step links time-series forecasting with probabilistic match outcome modeling, allowing direct comparison of competitive dynamics under projected player performance.

Figure 10 summarizes the logical pipeline used in this analysis. Forecasted Elo ratings from the ARIMA models serve as inputs to the classical Elo probability formulation, where the relative difference in expected performance levels determines the likelihood of victory.

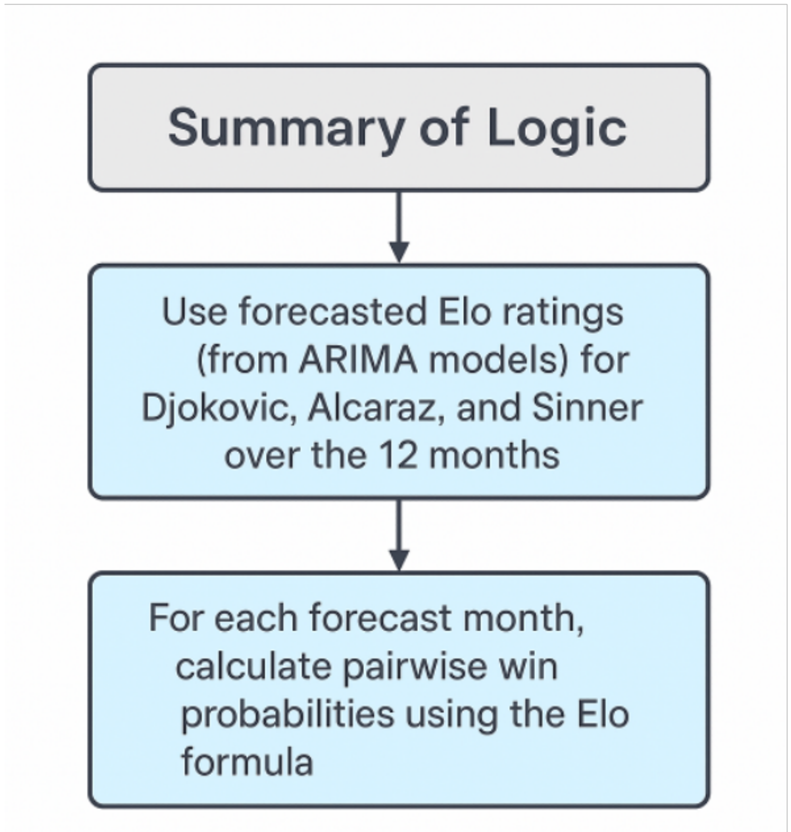

Figure 10: Workflow for converting forecasted Elo ratings into monthly win probabilities.

The pairwise probability between any two players, A and B, was derived using the standard Elo formula:

$$P(A \text{ beats } B) = \frac{1}{1 + 10^{-\frac{(E_A - E_B)}{400}}}$$

where E_A and E_B denote the forecasted Elo ratings of players A and B, respectively. This expression assumes a logistic mapping between Elo difference and expected win percentage, with a 400-point difference corresponding to approximately a 10:1 odds ratio.

Table 5: Next-month matchup probabilities based on forecasted Elo ratings.

Matchup	E_A	E_B	$P(A_{\textbf{wins}})$	$P(B_{\textbf{wins}})$	**Favorite**
Novak Djokovic vs Carlos Alcaraz	2056.91	1937.23	0.666	0.334	Novak Djokovic
Novak Djokovic vs Jannik Sinner	2056.91	2111.08	0.423	0.577	Jannik Sinner
Carlos Alcaraz vs Jannik Sinner	1937.23	2111.08	0.269	0.731	Jannik Sinner

Figures 11–13 display the month-ahead win probability trajectories for each key matchup. Djokovic maintains an advantage across both pairings, though the magnitude of this dominance varies over time. Against Alcaraz, his win probability stabilizes around 65–70%, while against Sinner it fluctuates more widely, reflecting the younger player's higher variance in performance projections. Meanwhile, the Alcaraz–Sinner matchup remains comparatively balanced, with alternating slight edges indicating their competitive parity in projected form.

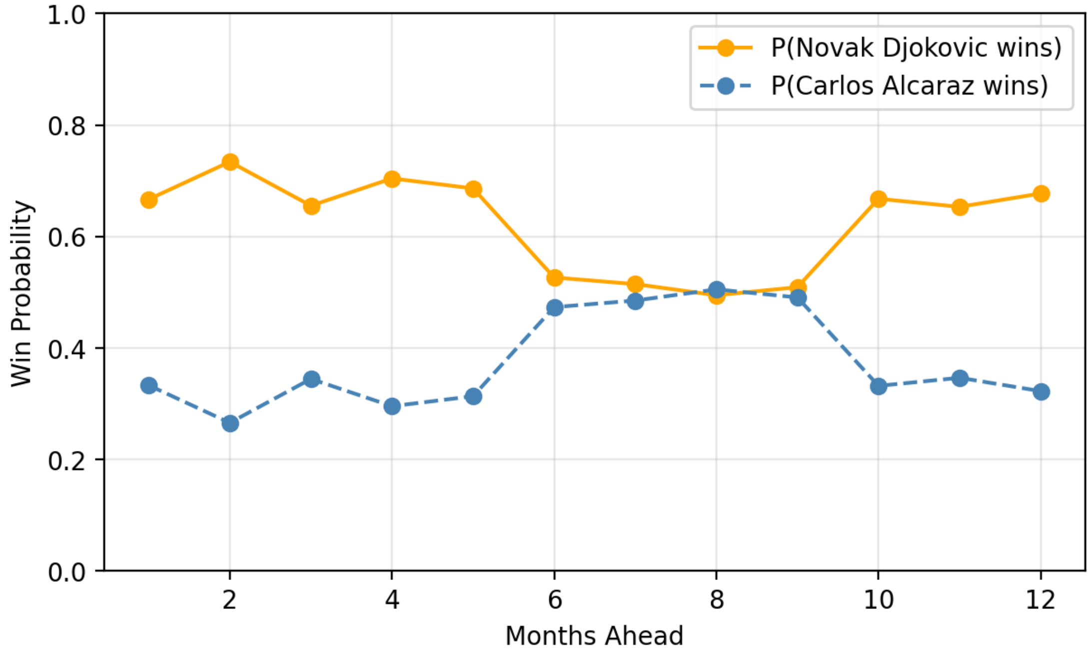

Figure 11: Projected monthly win probabilities for Novak Djokovic vs. Carlos Alcaraz.

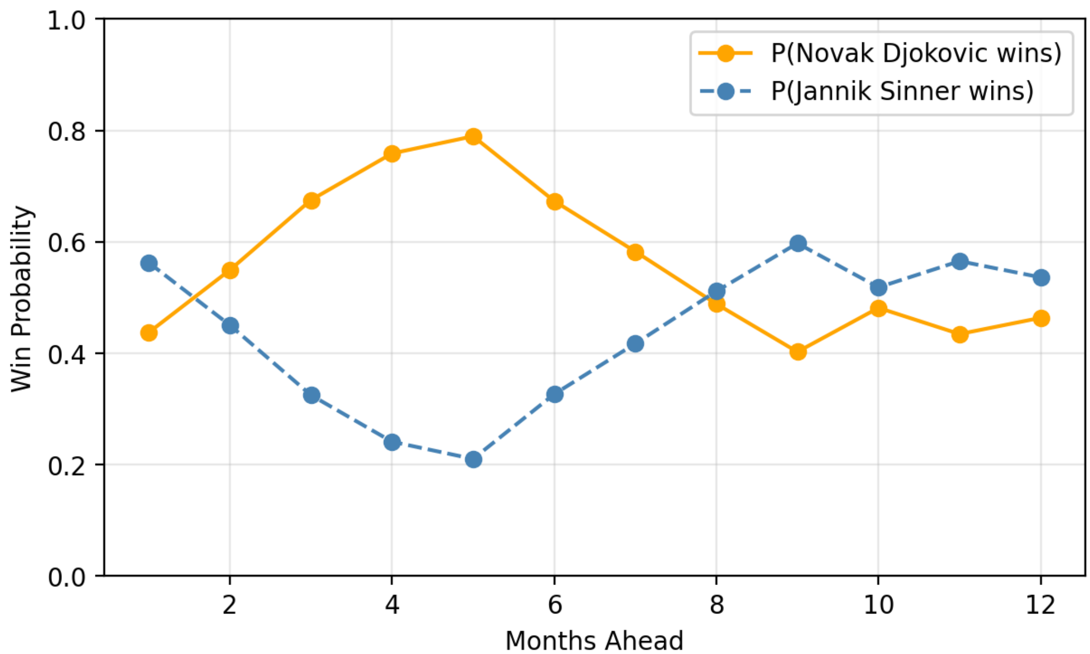

Figure 12: Projected monthly win probabilities for Novak Djokovic vs. Jannik Sinner.

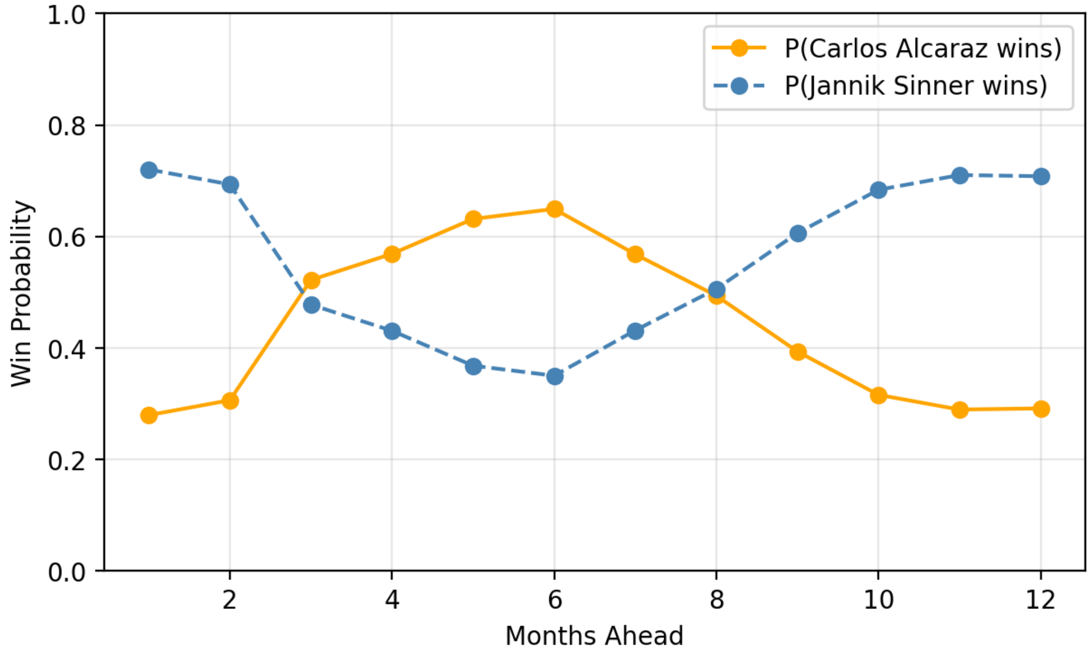

Figure 13: Projected monthly win probabilities for Carlos Alcaraz vs. Jannik Sinner.

7 Reflections and Lessons Learned

This project provided an end-to-end exploration of modeling player performance through Elo-based time series forecasting. While the ARIMA framework proved effective for capturing smooth rating trends, the process revealed several avenues for methodological and practical enhancement.

Modeling Improvements and Next Steps

SARIMAX emerged as the most natural extension of the current approach. It incorporates seasonal components and exogenous regressors—such as surface type, player fatigue, or injury indicators—allowing richer modeling of context-dependent variations. Its interpretability remains close to ARIMA while offering tangible gains in forecast realism and adaptability.

Hybrid ARIMA–LSTM (or ARIMA–XGBoost) frameworks represent a promising balance between statistical rigor and nonlinear learning capacity. By allowing ARIMA to model structured temporal dependencies and LSTM or XGBoost to capture residual nonlinearities, these hybrid architectures can improve predictive accuracy, particularly for players exhibiting abrupt or volatile Elo shifts across seasons.

State-Space and Kalman Filter models provide another frontier for refinement. These dynamic systems excel at real-time updating of player form after each match, combining smoothness with adaptability. While more computationally intensive and data-demanding, they hold strong potential for live forecasting scenarios where adaptability and interpretability must coexist.

Application Deployment and Broader Impact

An interactive Streamlit dashboard was developed to make the forecasting framework accessible beyond static notebooks. Users can select any ATP player and run the full pipeline—from Elo data extraction to ARIMA-based fitting, residual diagnostics, and 12-month forecasts with confidence intervals. The app generates real-time visualizations of Elo trends, residual behavior, and forecasted win probability trajectories, translating statistical output into intuitive insights for analysts, coaches, and fans. The Streamlit interface transformed the workflow into a reproducible and extensible research tool, enabling on-demand experimentation with modeling parameters.

References

[1] Sackmann, J. (2025). *ATP Tennis Data Repository*. Retrieved from https://github.com/JeffSackmann/tennis_atp

[2] Li, Z. (2025). *STA 9701: Time Series Analysis — Course Notes*. Baruch College, City University of New York.

Appendix: Repository Access

All project code, data processing scripts, and visualizations are available in the accompanying GitHub repository:

- `https://github.com/srijith-reddy/Elo-Based-Time-Series-Forecasting`

The repository includes the Jupyter notebook `Tennis Time Series.ipynb` and a Streamlit dashboard titled `ATP Elo Forecast Dashboard`. The app automates Elo computation, ARIMA-based forecasting, and visualization of key diagnostics and probability plots for any ATP player. A preview of the dashboard (`Elo_Dashboard.gif`) is included in the repository, along with setup instructions and dependencies in the `README.md` file.