Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD36544

Out-of-Sample Evaluation of Machine Learning for Macro-Factor-Based ETF Allocation

Jing Liu

School of Management and Economics, Beijing Institute of Technology, Beijing, China
jane1005lj@gmail.com

Abstract. Whether ETF dynamic allocation can use machine learning to transform macro information into effective stock-bond signals still requires rigorous out-of-sample testing. This paper uses SPY and TLT to represent U.S. equities and long-term Treasury assets, combined with Choice prices and FRED macro data, constructing a monthly sample of 235 periods from March 2006 to September 2025, and implementing an expanding-window forecast in the last 47 periods of final holdout samples. This paper compares the direction prediction ability of Logistic Regression, Random Forest, and XGBoost, and converts the probability of unweighted Logistic Regression output into the continuous allocation weights of stocks and bonds. The results show that each model does not stably exceed the simple benchmark, and the AUC confidence intervals of the two types of Logistic Regression cover 0.5; Under the moving-block bootstrap, the probability prediction error of the unweighted model is higher than the historical prevalence benchmark. The net cumulative return of the probability-weighted strategy is 10.42%, which is higher than the static 50/50 portfolio, but lower than the buy-and-hold SPY and the same-average-weight portfolio, and the confidence interval of the strategy return difference contains zero. The research shows that the allocation value of low-frequency macro and market characteristics is limited, and strict out-of-sample testing and asset exposure control help to identify the applicable boundaries of machine learning strategies.

Keywords: Machine learning, ETF asset allocation, Macro factors, Out-of-sample prediction, Probability-weighted allocation

1. Introduction

With the characteristics of low cost, transparent positions, and strong liquidity, exchange-traded funds (ETFs) have become an important tool for asset allocation. The return performance of stocks and long-term treasury bonds will be affected by multiple factors such as interest rate level, term structure, inflation, economic cycle, and investor risk appetite. Therefore, although the static stock–bond portfolios are simple to implement and provide diversification benefits, it is difficult to reflect the changes in the macro environment in time. Machine learning can handle high-dimensional features, nonlinear relationships, and variable interactions, offering new methods to transform macro information into dynamic allocation decisions. However, due to the high noise in financial markets and the constantly changing structural relationships, the historical fitting advantage may not be

DOI: 10.54254/2754-1169/2026.LD36544

sustained into the out-of-sample period. Therefore, machine learning allocation models must not only undergo predictive metric testing, but also evaluate their economic value under strict time constraints and strong benchmark comparisons. Previous studies have examined the application of machine learning in the financial field from multiple perspectives. Gu et al. found that the monthly out-of-sample R^2 values of random forests, gradient-boosted trees, and neural networks were about 0.33% -0.40%, which was higher than 0.26% -0.27% of the traditional linear method, indicating that the nonlinear relationship and variable interaction may provide incremental prediction information [1]. Pinelis further applied random forests to dynamic allocation between market indices and low-risk assets, providing direct evidence for the application of machine learning in asset allocation [2]. However, Kelly, Xiu, and Bagnara all pointed out that model flexibility also increases the risk of overfitting, interpretation difficulties, and out-of-sample instability [3, 4]. Zhang et al. found that the monthly alpha in the original report of a random forest strategy was 1.54%, but it basically disappeared after eliminating the forward-looking bias, indicating that the model performance may be highly sensitive to the information time alignment [5]. Traditional return predictability literature similarly indicates that most variables fail to persistently outperform historical benchmarks in extended samples [6-8], and complex combination methods may also fail to beat simple allocation rules due to estimation error [9]. Existing research mainly focuses on high-dimensional individual stocks, fund screening, or complex portfolio optimization. Whether low-frequency macro variables can support executable ETF stock-bond allocation, and whether dynamic strategies still generate incremental value after controlling average stock exposure, still requires further testing.

This paper uses SPY and TLT to represent the assets of U.S. stocks and long-term treasury bonds, respectively, constructs the characteristics of macro-level changes, ETF returns, momentum, and volatility, and compares the predicted performance of Logistic Regression, Random Forest, and XGBoost. The samples are sequentially divided into training, validation, and final holdout periods. Model selection is based on the training and validation phases, and the Logistic Regression parameter is determined only through internal time series cross-validation within the training period. The final holdout period uses an expanding-window forecast and maps the SPY positive return probability to the continuous equity-bond weight. The backtest also considers transaction costs, non-zero risk-free rates, static 50/50 portfolios, and the static same-average-weight portfolio benchmark. The main incremental aspect of this paper lies in strictly controlling information timing, isolating the final holdout sample, and distinguishing dynamic timing returns from long-term asset exposure returns using the same-average-weight benchmark.

2. Methodology

2.1. Data and variables

This paper uses SPY and TLT to represent U.S. stocks and long-term Treasury assets, respectively. The forward-adjusted closing prices come from Choice Financial Terminal, with the original range from January 2005 to December 2025; Each month, the simple monthly return is calculated using the price of the last valid trading day [10]. Macro and financial market data come from FRED, including long-term Treasury returns, term spreads, inflation, unemployment rate, industrial production, VIX, and effective federal funds rates [11]. The daily frequency variable uses the last valid observation at the end of the month, and the monthly frequency variable is aligned according to the corresponding month until the end of the month.

Candidate characteristics include the level and change of macro variables, the current month's return of SPY and TLT, the cumulative momentum and rolling volatility of 3 and 6 months. In order

DOI: 10.54254/2754-1169/2026.LD36544

to approximate the actual information availability time, CPI_YOY, UNRATE, and INDPRO_YOY are lagged by one month. After time alignment and missing value processing, the final sample covers 235 consecutive monthly observations from March 2006 to September 2025. Table 1 summarizes the definition, construction, and data sources of variables.

Momentum and volatility characteristics only use the returns realized at the end of the forecast time point and before. This paper sets the next SPY revenue direction as the target of two categories:

$$y_{t+1} = \begin{cases} 1, & R_{SPY,t+1} > 0, \\ 0, & R_{SPY,t+1} \leq 0. \end{cases} \tag{1}$$

Among them, the feature with date t only contains information available at the end of month t and earlier, used to predict the SPY returns direction for month $t+1$.

Table 1. Variable definitions, construction methods, and data sources

Variable group	Variables	Construction	Source
ETF returns	SPY_ret_1m, TLT_ret_1m	Compute simple monthly returns based on adjacent month-end backward-adjusted prices.	Choice
Interest-rate factors	DGS10, T10Y2Y, FEDFUNDS	Monthly level values and first-order changes	FRED
Macroeconomic factors	CPI_YOY, UNRATE, INDPRO_YOY	Monthly level values and first-order changes, uniformly lagged by one period	FRED
Market uncertainty	VIXCLS	The last valid observation of each month and its monthly change	FRED
Momentum factors	SPY/TLT 3- and 6-month momentum	As of forecast month t, calculate compound cumulative returns using the recent 3-month or 6-month monthly returns, including the current month.	Choice
Volatility factors	SPY/TLT 3- and 6-month volatility	As of forecast month t, calculate rolling standard deviations using the recent 3-month or 6-month monthly returns, including the current month	Choice
Prediction target	Next-month SPY return direction	Equals 1 if the next-period SPY return is strictly positive, and 0 otherwise	Choice

Note: Both momentum and volatility features are forecasted at month-end and include the monthly return at the forecast point. For example, the 3-month feature at the end of month t uses return information for months t, $t-1$ and $t-2$. Since the above returns were realized and observable at the end of month t, and the model predicts the direction of SPY returns for month t+1, this treatment does not include future information.

2.2. Candidate models and model selection

This paper sets the Always Positive Baseline and Historical Prevalence Baseline, and compares Logistic Regression, Random Forest, and XGBoost, representing linear classification, bagging-based tree, and boosting-based tree models, respectively. Model selection is mainly based on validation period AUC, while also checking whether the prediction has degenerated into the

DOI: 10.54254/2754-1169/2026.LD36544

majority class. The samples are divided into the training period, validation period, and final holdout period in chronological order, including 141, 47, and 47 observation periods, respectively. The final holdout period does not participate in model or feature selection.

The validation results did not show that Random Forest and XGBoost have stable increments, so eight core features with clear economic implications were ultimately adopted, while retaining the Logistic Regression with lower degrees of freedom. Balanced Logistic Regression is used for direction classification. During the training period, TimeSeries Split is used to select regularization strength and class weights based on AUC, ultimately setting $C = 0.01$ and $class\_weight = \prime balanced \prime$. Unweighted Logistic Regression used for probability allocation selects the regularization strength separately using the Brier Score, ultimately setting it to $C = 0.1$. All parameters are fixed before entering the final holdout period.

2.3. Final holdout sample prediction and allocation design

The final evaluation uses an expanding-window forecast method. Since November 2021, the model has re-estimated and predicted the next SPY positive return probability each month using only all previous observations, forming a total of 47 forecasts corresponding to the investment return period from December 2021 to October 2025. The design maintains chronological order and avoids using future information from the final holdout period [12]. Balanced Logistic Regression is used to calculate accuracy, Precision, Recall, F1-score, and AUC; Unweighted Logistic Regression is used to generate the allocation probability, and Brier score is used to evaluate the probability error. In this paper, an ex post selected probability threshold is not used, but the prediction probability is directly mapped to the SPY weight, and the TLT weight is its complement. This linear mapping makes the asset weight change smoothly with the prediction probability, reduces the weight jump and turnover caused by discrete switching, and directly reflects the model's confidence in the positive return of SPY. This setting retains probability information and enables continuous allocation without using leverage or short selling. Let p_t be the predicted positive return probability for the next SPY at the end of month t, then:

$$w_{SPY,t} = p_t, \quad w_{TLT,t} = 1 - p_t. \tag{2}$$

The backtest assumes a transaction cost of 10 basis points per dollar traded and calculates turnover based on post-return drifted weights. Strategy performance is evaluated using cumulative return, annualized return, annualized volatility, Sharpe Ratio, and maximum drawdown, with the risk-free rate converted from FEDFUNDS. To account for time dependence, a 6-month moving block bootstrap with 2,000 resamples is employed to estimate 95% confidence intervals for AUC, Brier Score, and monthly strategy return differentials.

Investment benchmarks include buy-and-hold SPY, a monthly rebalanced 50/50 SPY–TLT portfolio, and a monthly rebalanced static same-average-weight portfolio. The last benchmark uses the dynamic strategy's average SPY weight over the final holdout period and serves solely as an ex post diagnostic benchmark for separating returns attributable to asset exposure from those attributable to dynamic timing; it is not an ex ante implementable strategy.

DOI: 10.54254/2754-1169/2026.LD36544

3. Results

3.1. Candidate-model validation and final holdout results

Panel A of Table 2 reports the validation results of candidate models. As the proportion of positive returns in the month reached 70.2%, it is always predicted that the rise will constitute a strong benchmark. The AUCs of Logistic Regression, Random Forest, and XGBoost were 0.439, 0.316, and 0.461, respectively, and Random Forest classified 46 of the 47 validation observations as positive. The complex model did not show stable nonlinear increments.

Panel B and D of Table 2 report the final 47 retained sample results and confidence intervals. The AUCs of balanced and unweighted Logistic Regression were 0.488 and 0.451, respectively, and their 95% confidence intervals covered 0.5, which did not exhibit stable out-of-sample ranking ability. The Brier scores of Historical Prevalency Baseline and unweighted Logistic Regression were 0.233 and 0.246, respectively, with a difference of 0.013 and a 95% confidence interval of [0.0003, 0.0271]. The unweighted model produced a significantly higher Brier score than the historical prevalence benchmark under the adopted bootstrap procedure. In general, the selected macro and market characteristics do not provide stable out-of-sample forecast increments.

Table 2. Model evaluation and portfolio performance

Panel A. Validation-period candidate model performance

Model	Accuracy	Precision	Recall	F1-score	AUC
Naive Baseline Always Positive	0.702	0.702	1.000	0.825	0.500
Logistic Regression	0.660	0.707	0.879	0.784	0.439
Random Forest	0.681	0.696	0.970	0.810	0.316
XGBoost	0.553	0.658	0.758	0.704	0.461

Panel B. Final holdout classification performance

Model	Accuracy	Precision	Recall	F1-score	AUC	Brier Score
Always Positive Baseline	0.638	0.638	1.000	0.779	0.500	0.362
Historical Prevalence Baseline	0.638	0.638	1.000	0.779	0.291	0.233
Balanced Logistic Regression	0.447	0.643	0.300	0.409	0.488	0.260
Unweighted Logistic Regression	0.574	0.625	0.833	0.714	0.451	0.246

Panel C. Final holdout portfolio performance

Strategy	Cumulative Return	Annualized Return	Annualized Volatility	Sharpe Ratio	Max Drawdown
Probability-Weighted Strategy Gross	10.92%	2.68%	14.19%	−0.016	−27.90%

DOI: 10.54254/2754-1169/2026.LD36544

Panel C. (continued)

Probability-Weighted Strategy Net	10.42%	2.56%	14.20%	−0.024	−28.02%
Buy-and-hold SPY	58.37%	12.46%	16.23%	0.573	−23.93%
Static 50/50 SPY–TLT Gross	4.96%	1.24%	14.30%	−0.114	−26.81%
Static 50/50 SPY–TLT Net	4.88%	1.22%	14.30%	−0.116	−26.82%
Static Same Average Weight Gross	13.89%	3.38%	14.43%	0.034	−26.23%
Static Same Average Weight Net	13.80%	3.36%	14.43%	0.033	−26.24%

Panel D. Bootstrap confidence intervals for predictive metrics

Metric	Point Estimate	95% CI Lower	95% CI Upper
Balanced Logistic AUC	0.488	0.301	0.577
Unweighted Logistic AUC	0.451	0.274	0.564
Historical Prevalence Brier Score	0.233	0.197	0.276
Unweighted Logistic Brier Score	0.246	0.214	0.287
Brier Difference: Unweighted − Historical	0.013	0.0003	0.027

Panel E. Bootstrap confidence intervals for strategy return differences

Comparison	Mean Monthly Difference	95% CI Lower	95% CI Upper
Dynamic Net − Static 50/50 Net	0.109%	-0.072%	0.294%
Dynamic Net − Static Same Average Weight Net	-0.067%	-0.236%	0.083%

Table 2 Notes: Panel A reports candidate model performance during the validation period from December 2017 to October 2021; Panel B reports the classification results of the expanding-window forecast during the final holdout period from November 2021 to September 2025, with a lower Brier Score indicating smaller probability prediction errors; Panel C reports investment returns realized from December 2021 to October 2025. Net return is calculated at a cost of $10 basis points per transaction. Both the static 50/50 portfolio and the same-average-weight portfolio are rebalanced monthly; the same-average-weight portfolio serves as an ex post diagnostic benchmark for return attribution. Panel D reports 95% confidence intervals obtained from a six-month moving-block bootstrap with 2,000 resamples. Here, the Brier Score difference is defined as Unweighted Logistic Regression minus the Historical Prevalence Baseline, and a positive value means the unweighted model has a higher probability prediction error. Panel E reports the dynamic strategy's paired monthly return spread and its 95% confidence interval relative to two static portfolios.

3.2. Economic performance of allocation strategy

As shown in Panel C of Table 2, the probability-weighted strategy achieves an average SPY weight of 59.71% and an average monthly turnover of 9.64%. Net of transaction costs, its cumulative return is 10.42%, outperforming the static 50/50 portfolio (4.88%), but underperforming both the buy-and-

DOI: 10.54254/2754-1169/2026.LD36544

hold SPY strategy (58.37%) and the equal-average-weight benchmark (13.80%). The dynamic strategy returns a Sharpe ratio of −0.024 and a maximum drawdown of −28.02%, also failing to surpass the equal-average-weight benchmark.

Figure 1 plots the cumulative wealth of the four strategies. The horizontal axis represents the realized return month, while the vertical axis reports cumulative wealth normalized to one at the beginning of the holdout period. The blue, orange, green, and red lines represent the net probability-weighted strategy, buy-and-hold SPY, the net static 50/50 portfolio, and the net static same-average-weight portfolio, respectively.

Panel E of Table 2 shows that the average monthly return difference of the dynamic strategy relative to the 50/50 portfolio is 0.109%, relative to the same-average-weight portfolio is − 0.067%, and both 95% confidence intervals contain zero. Therefore, the current sample cannot provide robust evidence for the incremental returns of dynamic strategies relative to static benchmarks. The gross and net cumulative returns of the dynamic strategy are 10.92% and 10.42%, respectively, indicating that the transaction cost is not the main reason for insufficient performance.

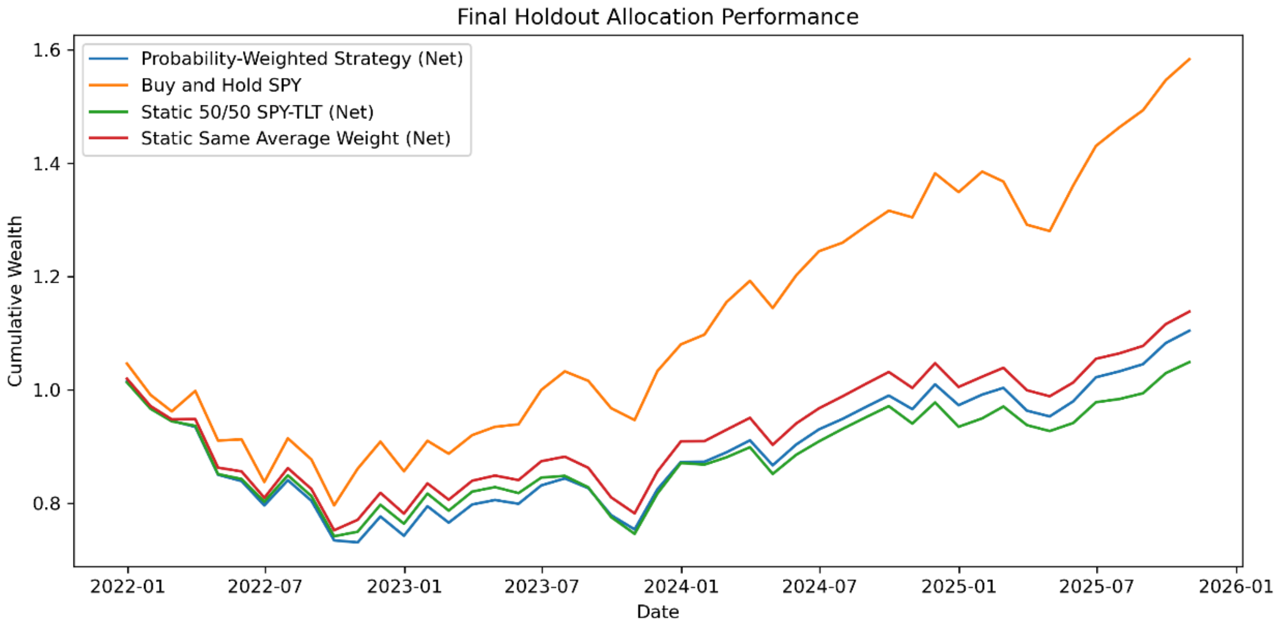

Figure 1. Out-of-sample cumulative wealth (photo/picture credit: original)

4. Discussion

This paper finds that the low-frequency macro and market characteristics fail to form a stable monthly return direction signal, and the complex tree model is not consistently superior to the linear model. This is consistent with the risk of out-of-sample failure, parameter instability, and over fitting highlighted in existing studies [6-8, 12]. The results of this paper do not mean that machine learning is generally ineffective in finance, but show that the complexity of the model can be transformed into stable value outside the sample only when the data size, signal strength, and information structure are sufficient to support it. Compared with the more positive results of machine learning in the fund cross-section screening and price graphic feature scenarios [13, 14], the low-frequency macro dual asset environment in this paper contains lower information dimensions and more limited prediction space.

DOI: 10.54254/2754-1169/2026.LD36544

The limited prediction ability may be related to data frequency, sample size, and goal setting. There are only 235 monthly observations in this paper, and the initial candidate features reach 24. The change of macro variables is slow, and its impact may take a longer time to transmit, while the monthly return is also impacted by policies, events, emotions, and liquidity. The positive return months of SPY account for the majority, which also makes the prediction of rising all the time become a strong benchmark. More importantly, predicting "whether SPY will rise" is not the same as predicting "whether SPY is better than TLT", so there is a target mismatch between the direction classification and the relative allocation of stocks and bonds.

High inflation and monetary tightening in 2022 further weakened the traditional stock-bond diversification mechanism. The Bank for International Settlements pointed out that in a high-inflation environment, tightening monetary conditions may lead to simultaneous declines in stocks and bonds [15]. TLT is sensitive to long-term interest rates, so even if the model reduces its SPY allocation, it may not be able to reduce risk by increasing its holdings in TLT. Figure 1 shows that the main drawdowns of the three types of stock-bond portfolios were concentrated during this period, and dynamic adjustments did not provide additional tail protection.

From an economic value perspective, the point-estimate advantage of the probability-weighted strategy over the 50/50 portfolio is insufficient to demonstrate timing efficacy. The dynamic strategy exhibits a higher average equity weight; once controlling for average SPY exposure, its return and risk metrics fail to outperform the static diagnostic benchmark, with the confidence interval for return differentials spanning zero. Therefore, this paper finds no evidence that dynamic adjustment creates robust incremental economic value. The limitations of this paper include the short final holdout period, the possible subsequent revision of FRED data, the limited scope of assets to SPY and TLT, and the simplified setting of transaction costs and market friction. Future research can use ALFRED real-time data to directly predict the relative return of SPY versus TLT, and expand to multi-asset allocation such as gold, credit bonds, overseas stocks, and cash [16].

5. Conclusion

Based on 235 monthly data points from March 2006 to September 2025, this paper compares whether Logistic Regression, Random Forest, and XGBoost can use macro and market characteristics to predict the next SPY return direction, and implements expanding-window allocation in the last 47 independent holdout samples. Results show that the model does not consistently exceed the simple prediction benchmark; the AUC confidence intervals for both logistic specifications included 0.5, and the unweighted model's probability error is also higher than the historical proportional benchmark. The net cumulative return of the probability-weighted strategy is 10.42%, estimated to be higher than the static 50/50 portfolio but lower than Buy-and-Hold SPY and the same-average-weighted portfolio, with both confidence intervals for the return differentials including zero. Overall, current low-frequency macro dual-asset scenarios do not show stable machine learning predictions or timing values, but strict out-of-sample isolation, statistical intervals, and asset exposure controls help identify the true sources of strategy performance.

References

[1] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[2] Pinelis, M. (2022). Machine learning portfolio allocation. *The Journal of Finance and Data Science, 8*, 35–54. https://doi.org/10.1016/j.jfds.2021.12.001

[3] Kelly, B. T., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance, 13*(3–4), 205–363. https://doi.org/10.1561/0500000064
[4] Bagnara, M. (2024). Asset pricing and machine learning: A critical review. *Journal of Economic Surveys, 38*(1), 27–56. https://doi.org/10.1111/joes.12532
[5] Zhang, Y., Zhu, Y., & Linnainmaa, J. T. (2025). Man versus machine learning revisited. *The Review of Financial Studies, 38*(12), 3768–3790. https://doi.org/10.1093/rfs/hhaf066
[6] Goyal, A., Welch, I., & Zafirov, A. (2024). A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 37*(11), 3490–3557. https://doi.org/10.1093/rfs/hhae044
[7] Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 21*(4), 1455–1508. https://doi.org/10.1093/rfs/hhm014
[8] Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies, 21*(4), 1509–1531. https://doi.org/10.1093/rfs/hhm055
[9] DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies, 22*(5), 1915–1953. https://doi.org/10.1093/rfs/hhm075
[10] East Money Information Co., Ltd. (2013). Choice Financial Terminal [DB/OL]. https://choice.eastmoney.com/
[11] Federal Reserve Bank of St. Louis. (1991). Federal Reserve Economic Data (FRED) [DB/OL]. https://fred.stlouisfed.org/
[12] Barroso, P., & Saxena, K. (2022). Lest we forget: Learn from out-of-sample forecast errors when optimizing portfolios. *The Review of Financial Studies, 35*(3), 1222–1278. https://doi.org/10.1093/rfs/hhab041
[13] DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & Santos, A. A. P. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics, 150*(3), 103737. https://doi.org/10.1016/j.jfineco.2023.103737
[14] Murray, S., Xia, Y., & Xiao, H. (2024). Charting by machines. *Journal of Financial Economics, 153*, 103791. https://doi.org/10.1016/j.jfineco.2024.103791
[15] Bank for International Settlements. (2022). Old challenges, new shocks. In *BIS Annual Economic Report 2022* (Chapter I). Bank for International Settlements. https://www.bis.org/publ/arpdf/ar2022e1.htm
[16] Jensen, T. I., Kelly, B., Malamud, S., & Pedersen, L. H. (2026). Machine learning and the implementable efficient frontier. *The Review of Financial Studies*, hhag022. https://doi.org/10.1093/rfs/hhag022