

Meral KAGITCI, PhD (corresponding author)

meral.kagitci@ase.ro

Bucharest University of Economic Studies, Bucharest, Romania

Şener ALI, PhD Student

alisener20@stud.ase.ro

Bucharest University of Economic Studies, Bucharest, Romania

Stock Portfolio Optimization with Explainable AI: A Comparative Study with the Black-Litterman Model

Abstract. *This study aims to determine whether an artificial intelligence (AI) approach can outperform Black-Litterman (BL) - a classical stock portfolio optimization framework. Using a diversified dataset that contains market indices, sector exchange-traded funds (ETFs) and stocks, we compare the performance of XGBoost with a Black-Litterman benchmark in terms of returns, risk, drawdowns, and turnover. Furthermore, we evaluate the economic interpretability of Shapley Additive exPlanations (SHAP)-based explanations. The BL approach achieved a higher annualized return, whereas XGBoost obtained lower volatility, a shallower maximum drawdown, and substantially lower turnover. SHAP diagnostics show that the AI model's decisions are largely driven by trend and volatility indicators, providing explanations that are faithful to the model and aligned with financial intuition. The findings demonstrate that explainable AI (XAI) can enhance transparency and help investors refine their views on expected returns and risks, but that robust, economically grounded frameworks like Black-Litterman remain essential as anchors for risk-controlled portfolio construction.*

Keywords: *stock portfolio optimization, portfolio risk management, Black-Litterman model, explainable artificial intelligence.*

JEL Classification: C45, G11, G17.

Received: 11 December 2025	Revised: 14 September 2026	Accepted: 21 September 2026
-----------------------------------	-----------------------------------	------------------------------------

1. Introduction

Technological advancement has made access to global financial markets easier for all investors. Whether it is a micro-investor or an important investment fund, the overwhelming number of choices, the dynamic environment, and the large amount of available data – news, macroeconomic indicators, sentiment measures, analysts' recommendations – can significantly increase the complexity of managing a stock portfolio. Investors therefore need tools capable of addressing these challenges (Freibauer et al., 2024).

Over time, investors have developed multiple frameworks for turning views about risk and return into asset allocations. The Black-Litterman model represents

one of the most influential of them (Black and Litterman, 1992; Satchell and Scowcroft, 2000). In today's market environment, the model presents several limitations, such as unstable correlations across regimes and higher liquidity/transaction-cost frictions (Sun et al., 2023). Despite these limitations, it remains a reliable and a widely used tool for portfolio optimization (Chincarini and Kim, 2013).

Recent advances in artificial intelligence (AI), especially in deep learning (DL), have made it possible to build algorithms capable of learning complex, nonlinear relationships, which are well-suited for forecasting stock prices. Despite their ability to process large amounts of data, AI models remain limited by a lack of transparency, which may constrain their adoption by asset managers. In practice, investors need to understand the reasons underlying model-generated decisions. Even when a model demonstrates superior performance, insufficient explainability may reduce portfolio managers' trust (Fischer and Krauss, 2018).

XAI provides tools to mitigate this opacity. We can determine which features drive the model's forecasts by integrating methods such as SHAP values or integrated gradients. However, such explanations may appear intuitive without necessarily being economically meaningful. To be more than a visualization layer, we need to evaluate the results through tests of fidelity and decision impact (Lundberg and Lee, 2017).

This article aims to investigate whether an AI-driven portfolio-optimization strategy can outperform classical baselines, and whether greater transparency and control can be achieved by using XAI. In this way, our study is guided by the following research questions:

RQ1. Can an AI model consistently outperform the classical Black-Litterman model, while remaining robust across market regimes and stress periods?

RQ2. Do XAI methods produce explanations that are faithful to the model and economically meaningful?

The remainder of the paper is structured as follows. Section 2 reviews the related literature, Section 3 describes the data, Section 4 presents the methodology, Section 5 reports and discusses the results, and Section 6 concludes the study and outlines directions for future research.

2. Literature Review

Stock market portfolio optimization went through multiple phases over the years, starting with the mean-variance paradigm developed by Harry Markowitz to more advanced methods such as the Black-Litterman model (Kolm et al., 2014). Introduced in the early 1990s, the BL model starts from an expected performance based on market data and then updates those expected returns by treating the investors' views as noisy observations about returns (Kolm and Ritter, 2017).

Modern research extends the BL model to better match asymmetry in returns. One approach proposes a stable inverse-optimization procedure for portfolio construction by extracting market information and blending this information with

structured views. Empirical tests showed that the extension to non-Gaussian distributions performed better than the classical model (Palczewski and Palczewski, 2019). Similarly, a copula-based BL variant reports improvements in risk and risk-adjusted performance versus benchmarks (Sahamkhadam et al., 2022).

A dynamic BL model that incorporates time variation and tail risk has also demonstrated improved performance (Harris et al., 2017). Furthermore, the BL model's handling of nonlinearities was optimized by combining it with neural forecasting methods for views (Simos et al., 2021). Finally, Jayadi & Sumarti (2023) obtained better outcomes by generating investor views using a Long Short-Term Memory (LSTM) model.

Beyond traditional forecasting methods for stock portfolio optimization, the literature has increasingly adopted more advanced techniques, including machine learning (ML) and deep learning (DL) models, which have demonstrated improved performance (Ma et al., 2021). Researchers have explored several ways to improve AI models. Ashrafzadeh et al. (2025) improved algorithmic performance through clustering methods (K-means, Dynamic Time Warping (DTW), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)), improving risk control by integrating a constant absolute risk aversion (CARA) utility function achieved better results when tested on S&P 500 data (Kubo and Nakagawa, 2025).

Despite their improved performance in forecasting, DL models are also used for selection and ranking. The combination of Deep RankNet models and genetic-algorithm hyperparameter optimization demonstrated strong performance even in volatile periods such as COVID-19 pandemic (Alzaman, 2024). Meanwhile, the rise of deep reinforcement learning (DRL) reframes portfolio choice, providing more efficient and profitable strategies (Espiga-Fernández et al., 2024).

DRL models' abilities can be enhanced through various methods. Soleymani and Paquet (2021) combined DRL with graph convolutional networks (GCNs) to model dynamic relationships among financial assets and obtained robust results even in volatile periods. Similarly, an actor-critic design argues that combining local asset embeddings with a learned global context can enhance DRL's capabilities (Yang, 2023). Furthermore, Jang and Seong (2023) combined DRL with Modern Portfolio Theory (MPT) and reported favourable results for the Dow Jones index.

Parallel to this performance-driven shift toward ML/DL portfolio construction, the literature draws attention to the importance of transparency. It is essential for investment managers and traders to understand why a model recommends buying or selling a stock. Black-box opacity therefore represents a key barrier to operational use in real investment workflows (Sutiene et al., 2024). The level of transparency can be enhanced using XAI methods, such as SHAP and Local Interpretable Model-Agnostic Explanations (LIME) (Černevičienė and Kabašinskas, 2024).

Weber et al. (2024) showed that interest in XAI has increased substantially because these methods can improve model transparency. The explanations provided by the model should be adapted to the level of understanding of the end user (Yeo et al., 2025). Furthermore, it is imperative to test how economically significant the

explanations are to establish a meaningful link between model interpretability and asset-pricing relevance (Goswami and Uddin, 2025).

Taken together, the literature can be grouped into three main research streams: extensions of the Black-Litterman framework, ML/DL-based portfolio construction, and XAI approaches. The closest studies to our setting combine Black-Litterman with model-generated views, including neural forecasting approaches and the LSTM-based approach. In contrast, our study directly compares an XGBoost-based allocation strategy with a Black-Litterman benchmark over a common out-of-sample period and complements this comparison with SHAP-based analysis of the XGBoost model.

3. Data

This section describes the process of building the dataset used for both components of the study, namely the estimation of Black-Litterman model outputs and the training and evaluation of the XAI-based model.

3.1 Data Fetching and Sampling Design

The data used in this study were collected from Yahoo Finance through the *yfinance* library. We retrieved daily observations for the period from 1 January 2000 to 1 December 2025, including open, close, high, low, and volume. The collected data were subsequently stored in a local SQLite database using a Python script. Our database is designed to be incrementally updatable over time and usable for both classical and XAI models.

3.2 Selected Stocks

To support a broad comparison, we selected a broad set of market indices, sector ETFs and large-cap stocks. This reduces the dependence on any single sector when comparing BL and XAI models under different correlation and volatility regimes. Because some of the selected assets were listed after 2000, observations for these assets are available only from their respective first trading dates. Table 1 presents the financial assets included in the database.

Table 1. Financial ASSETS Available in the Database

Tickers	Group
GSPC, NDX, SPY, QQQ, IWM	Market benchmarks
XLK, XLF, XLE, XLV, XLI, XLP, XLY, XLU, XLB, XLRE, XLC	Sector ETFs
AAPL, MSFT, NVDA, INTC, CSCO, ORCL, IBM, ADBE, QCOM, TXN, AVGO, AMD, MU, AMAT, LRCX, KLAC, ADI	Technology – AI Stack
AMZN, GOOGL, META, NFLX, TSLA, CRM	Technology
BRK-B, JPM, BAC, WFC, C, GS, MS, AXP	Financials

Tickers	Group
JNJ, PFE, MRK, UNH, ABBV, ABT, TMO, AMGN, GILD	Healthcare
PG, KO, PEP, COST, WMT	Consumer staples
HD, MCD, NKE, DIS	Consumer discretionary
CAT, HON, GE, BA, UNP, UPS, LMT	Industrials/defense/transport
XOM, CVX, COP, SLB	Energy
LIN, APD, NEM	Materials
NEE, DUK, SO, PLD, SPG, O	Utilities + REITs
VZ, T, TMUS, CMCSA	Communications

Source: Authors' own creation.

3.3 Database Design

The database is a file-based SQLite store. Each data type is stored in a normalized structure to ensure traceability and avoid duplication. Table 2 presents the database tables and their description.

Table 2. Database Tables

Table	Granularity	Description
tickers	Per ticker	Metadata snapshot (name, exchange, currency, sector/industry, country)
prices	Per ticker-day	Open – high – low – close prices, volume
actions	Per corporate action event	Dividends and stock splits
info_json	Per ticker snapshot	Raw JSON info from Yahoo Finance
features	Per ticker-day	Engineered indicators

Source: Authors' own creation.

3.4 Feature Construction (for XAI Training)

To achieve the XAI objective of this study, we construct a set of technical and volatility indicators based on open, high, low, close and volume data at the date-ticker level. Short-horizon returns (`ret_1`, `log_ret_1`) capture immediate price moves, while a set of simple moving averages over different windows (`sma_10`, `sma_20`, `sma_50`) encode short- and medium-term trend structure. Realized risk is represented by the 20-day annualized volatility (`ann_vol_20`) and the 14-day Average True Range (`atr_14`), which jointly describe both return variability and intraday trading ranges. Volume-related features (`vol_sma_20`, `vol_ratio`) provide a proxy for liquidity and participation, allowing the model to condition its forecasts on whether price moves occur in high- or low-volume environments.

To complement these basic building blocks, we include standard oscillator-type indicators that are widely used in practice to characterize momentum and overbought/oversold conditions. The 14-day Relative Strength Index (`rsi_14`) measures recent upside versus downside pressure, while the Moving Average Convergence Divergence (MACD) indicator (`macd`, `macd_signal`, `macd_hist`) summarizes the interaction between short- and longer-term moving averages in a

more flexible way. Together, these features give the model access to information about trend, mean reversion, volatility regimes and liquidity without forcing it to learn long-term dependencies from raw prices.

4. Methodology

This section presents the complete workflow used to compare the BL benchmark with the XGBoost-based portfolio strategy. The procedure consists of five sequential stages. First, daily market observations are cleaned, duplicate intraday timestamps referring to the same trading day are collapsed to a single day observation, and non-investable index and yield series are excluded, resulting in a common list of 87 investable assets. Second, the BL portfolio is estimated using rolling historical information and monthly rebalancing. Third, XGBoost is trained using a chronological, leakage-controlled design and is subsequently used to rank assets at each month-end rebalancing date. Fourth, the two strategies are evaluated over the same out-of-sample period using return, volatility, Sharpe ratio, maximum drawdown (MDD), Calmar ratio, and turnover. Finally, SHAP values are used to examine the contribution of the technical predictors to the XGBoost forecasts. The following subsections describe each stage in detail.

Black-Litterman was selected as the classical benchmark because it represents an established portfolio-allocation framework that explicitly incorporates expected returns and risk, whereas XGBoost provides a data-driven nonlinear alternative. The comparison is therefore intended to contrast two distinct portfolio-construction paradigms.

4.1 Black-Litterman Benchmark

The BL implementation follows the historical-allocation framework introduced by Black and Litterman (1992). Historical observations from January 2000 to December 2025 are retained for model estimation. Before estimation, duplicate database timestamps referring to the same calendar trading day are collapsed to one daily observation. Four non-investable series (GSPC, NDX, VIX, and IRX) are excluded from the portfolio universe, leaving 87 investable stocks and exchange-traded funds.

To estimate covariance matrices and views, we use a rolling window of length $N = 3 \times 252 = 756$, which represents approximately three years of daily trading data. Covariance matrices are estimated using pairwise observations with a minimum of 60 overlapping returns. To reduce estimation instability, the sample covariance matrix is shrunk toward its diagonal, and a regularization term is added when necessary. Portfolio weights are constrained to be long-only and to sum to one.

For the Black-Litterman strategy, we build the covariance matrix R_t with size $N \times n$, representing the daily returns for the last N days and n assets. Missing values are handled conservatively: diagonal NaNs are replaced by the average of the finite diagonal elements, and off-diagonal NaNs are set to zero.

We approximate the market equilibrium portfolio (ω_i^{mkt}) by equal weights across all assets, which serves as a practical proxy for a diversified market portfolio in the absence of long-history capitalization data:

$$\omega_i^{mkt} = \frac{1}{n}, i = 1, \dots, n \quad (1)$$

After that, we calculate the second essential component of the Black-Litterman model, the data-driven views, based on sample average returns over the same three-year window, using the following formula:

$$\mu_i^{sample} = \frac{1}{N} \sum_{t=1}^N r_{t,i} \quad (2)$$

All quantities are expressed in daily units. The BL posterior combines an equal-weight equilibrium prior based on the regularized covariance matrix and a risk-aversion coefficient of $\delta = 3.0$ with one absolute view per asset. The view matrix is $P = I$, the view vector Q contains the rolling 756-day mean returns, and prior uncertainty is $\pi = 0.025$. View uncertainty is represented by a diagonal matrix whose elements are the corresponding variances from the regularized covariance matrix after scaling by π . Given μ^{BL} and Σ in daily units, we compute the long-only maximum-Sharpe portfolio by solving:

$$\omega^* = \underset{\omega}{argmax} \frac{\omega^T \mu^{BL}}{\sqrt{\omega^T \Sigma \omega}} \quad (3)$$

with a zero risk-free rate for simplicity. The optimization is implemented with a Sequential Least Squares Programming (SLSQP) solver (SciPy) and falls back to a heuristic Sharpe-ratio tilt or equal weights if the numerical solution fails or yields non-finite weights.

The resulting optimal weights are recomputed at each month-end date using only information available up to that date. Between two rebalancing dates, the most recently determined weights are held constant, and realized portfolio returns are calculated using lagged weights so that a portfolio formed at date t affects returns only from the following trading day.

4.2 XGBoost Implementation

The XGBoost implementation applies the same data-cleaning and missing-value procedures as the BL benchmark. Based on these cleaned prices, we construct daily log prices using the following formula:

$$L_{t,i} = \ln P_{t,i} \quad (4)$$

which are used for prediction target definition and realized portfolio returns computation.

The prediction target is defined for a fixed horizon of $H = 63$ trading days (approximately three months). To ensure that the model is trained only on fully realized forward returns, we discard observations for which the corresponding future return cannot be computed.

The features table of the database provides 13 technical indicators – short-term returns (ret_1 , log_ret_1), volatility measures (ann_vol_20 , atr_14), simple moving averages (sma_10 , sma_20 , sma_50), oscillator indicators (rsi_14 , macd , macd_signal , macd_hist), and volume-related variables (vol_sma_20 , vol_ratio) – that we use as predictors. They are indexed by timestamp, resulting in a panel dataset:

$$\mathcal{D} = \{(x_{t,i}, y_{t,i})\} \quad (5)$$

where each observation corresponds to one asset on one date.

The XGBoost implementation follows Chen and Guestrin (2016) and uses a strictly chronological holdout design rather than random sampling. The out-of-sample evaluation boundary is set as 31 January 2018. Because the prediction target is the forward return over $H = 63$ trading days, observations whose target horizon could reach the evaluation period are excluded from model training. This separation prevents future evaluation-period returns from entering model estimation.

The predictive model is represented by an XGBoost regressor set up to balance flexibility and robustness. Table 3 presents the parameters and their values that we used for configuring the model.

Table 3. XGBoost Parameters

Parameter	Value	Description
n_estimators	400	Number of boosting trees (iterations)
max_depth	4	Maximum depth of each tree (shallower trees to reduce overfitting)
learning_rate	0.03	Shrinkage factor for each tree's contribution
subsample	0.7	Fraction of training samples used for tree
colsample_bytree	0.7	Fraction of features used per tree
reg_lambda	1.0	L2 regularization strength on weights
reg_alpha	0.0	L1 regularization strength on weights
min_child_weight	10	Minimum sum of instance weights in a leaf
objective	"reg:squarederror"	Regression with squared error loss
tree_method	"hist"	Histogram-based tree construction
random_state	42	Random seed

Source: Authors' own creation.

The fitted XGBoost model is used at each month-end to predict the 63-trading-day forward return for every eligible asset. Assets are ranked according to these predictions and the upper 50% of ranked assets form the subset used for the XGBoost-driven allocation. The final allocation combines an 80% equal-weight diversified base across eligible assets with a 20% XGBoost tilt allocated equally

among the selected assets. All pre-scaling weights are non-negative and normalized to sum to one. A 20% volatility target is subsequently applied using a 63-trading-day realized volatility window. The scaling factor is capped at one so that the strategy can reduce exposure without employing leverage. Portfolio returns are calculated using one-day-lagged weights.

After model training, we construct the stock portfolio by computing daily log returns for all assets over the entire price history and retaining only the subset of tickers that appear in the merged feature-target panel. Next, we construct a sequence of month-end rebalancing dates that fall strictly after the training-test split date. We then extract the cross-section of feature vectors for all assets on that feature date, reindex it to exactly match the training feature set (same columns in the same order), and fill any missing values with the pre-computed training medians.

4.3 Explainability via SHAP

To assess model interpretability, SHAP values are calculated following Lundberg and Lee (2017), on a reproducible random sample of up to 5000 training observations. Global feature importance is measured as the mean absolute SHAP value of each predictor rather than its signed average contribution. The resulting importances are additionally grouped into five economic categories (trend, volatility, volume, oscillators, and short-term returns) to facilitate interpretation of the technical signals underlying the XGBoost forecasts.

4.4 Performance Evaluation

This subsection defines the performance metrics used to evaluate and compare the two implemented models. For each model, we measure six key metrics as shown in Table 4. Annualized return expresses portfolio performance on an annual basis, while annualized volatility captures the variability of returns. The Sharpe ratio measures return relative to volatility, MDD represents the largest peak-to-trough loss, the Calmar ratio relates annualized return to maximum drawdown, and turnover captures the intensity of portfolio reallocation.

Table 4. Metrics

Metric	Symbol	Formula
Annualized return	$\mathcal{R}_{ann}$	$\left(\prod_{t=1}^T (1 + \mathcal{R}_t^{(p)}) \right)^{\frac{252}{T}} - 1$
Annualized volatility	σ_{ann}	$\sqrt{252} \text{sd}(\mathcal{R}_t^{(p)})$

Metric	Symbol	Formula
Sharpe ratio	Sharpe	$\frac{\mathcal{R}_{ann}}{\sigma_{ann}}$
Maximum drawdown	MDD	$\min_t \left(\frac{w^t}{\max_{s \leq t} W_s} - 1 \right),$ $W_t = \prod_{u \leq t} (1 + \mathcal{R}_u^{(p)})$
Calmar ratio	Calmar	$\frac{\mathcal{R}_{ann}}{ MDD }$
Turnover	τ	$\frac{1}{2} \sum_i \omega_{t_k, i} - \omega_{t_{k-1}, i} $

Source: Authors' own creation.

4.5 Out-of-Sample Validation and Robustness Analysis

The predictive evaluation follows a chronological holdout design, which is appropriate when estimating performance on temporally ordered financial observations (Cerqueira et al., 2020). The XGBoost model is estimated exclusively using observations available before the out-of-sample boundary. Given the 63-trading-day forecast horizon, the final training feature observation is dated 27 October 2017 and its corresponding target is fully realized by 30 January 2018. Portfolio formation begins on 31 January 2018, while realized returns for both strategies are compared over the common period from 1 February 2018 to 28 November 2025 and over the same set of 87 investable assets.

Two additional procedures are used to assess robustness. First, the common out-of-sample return series is divided into three economically distinct temporal subperiods: 2018-2019, 2020-2021, and 2022-2025. The same performance measures are recalculated within each subperiod without changing the baseline model specifications. Second, statistical uncertainty surrounding the differences between BL and XGBoost is assessed using a paired moving-block bootstrap, which preserves short-range temporal dependence by resampling observations (Künsch, 1989). We use 5000 bootstrap replications with blocks of 21 trading days. For each replication, paired BL and XGBoost observations from the same dates are resampled jointly, and the differences in the calculated metrics are recomputed. Statistical

significance is evaluated using two-sided bootstrap p-values and 95% confidence intervals, with a 5% significance threshold.

5. Results

This section presents the empirical results obtained by the Black-Litterman and XGBoost strategies and their economic interpretation.

5.1 Comparative Analysis

Both strategies are evaluated using the same 87 investable assets. Portfolio formation begins on 31 January 2018, and the common realized-return series extends from 1 February 2018 to 28 November 2025. Table 5 reports the performance of the two strategies.

Table 5. Results

Metric	Black-Litterman	XGBoost
Annualized return	0.1317 (13.17%)	0.1009 (10.09%)
Annualized volatility	0.2062 (20.62%)	0.1596 (15.96%)
Sharpe ratio	0.6387	0.6320
Maximum drawdown	-0.3447 (-34.47%)	-0.2687 (-26.87%)
Calmar ratio	0.3822	0.3754
Turnover	2.0831 (208.31% of portfolio/year)	0.5061 (50.61% of portfolio/year)

Source: Authors' own creation.

The results reveal a trade-off rather than uniform dominance by either strategy. BL achieves a higher annualized return (13.17% versus 10.09%), but this is accompanied by higher annualized volatility (20.62% versus 15.96%) and a deeper maximum drawdown (-34.47% versus -26.87%). The two strategies exhibit almost identical risk-adjusted performance. Their Sharpe ratios are 0.6387 and 0.6320, respectively, while their Calmar ratios are 0.3822 and 0.3754. A particularly large difference appears in trading intensity, with annual turnover of 208.31% for BL compared with 50.61% for XGBoost. Thus, BL generates a higher gross return, whereas XGBoost produces a lower-risk and substantially lower-turnover allocation.

Transaction costs are not explicitly modelled. Given BL's substantially higher turnover, incorporating such costs could reduce its observed gross-return advantage. Whether these descriptive differences are statistically significant is examined in the following section. Figure 1 illustrates the cumulative out-of-sample performance of the two strategies over the common evaluation period.



Figure 1. Cumulative performance of BL and XGBoost

Source: Authors' own creation.

5.2 Robustness and Statistical Significance

The temporal robustness analysis indicates that the relative performance of the two strategies varies across market periods. In 2018-2019, BL slightly exceeds XGBoost in annualized return (11.30% versus 9.97%) and Sharpe ratio (0.731 versus 0.673). The difference is substantially larger during 2020-2021, when BL achieves an annualized return of 36.69% and a Sharpe ratio of 1.292, compared with 15.26% and 0.847 for XGBoost, although BL also exhibits considerably higher volatility. In contrast, the ranking reverses during 2022-2025. XGBoost achieves a higher annualized return (7.57% versus 3.54%), a higher Sharpe ratio (0.492 versus 0.198), and a shallower maximum drawdown (-23.94% versus -34.43%). The complete subperiod results are reported in Appendix Table A1. These findings indicate that performance is regime-dependent and do not support the conclusion that either strategy dominates consistently across the full evaluation period. Table 6 presents the paired bootstrap results for the performance differences between the BL and XGBoost strategies.

Table 6. Paired Bootstrap Comparison

Metric BL - XGBoost	Difference	95% bootstrap CI	p-value	Significant at 5%
Annualized return	3.09%	[-6.55%, 13.51%]	0.5608	No
Annualized volatility	4.67%	[2.52%, 7.35%]	<0.001	Yes
Sharpe ratio	0.0067	[-0.5410, 0.5076]	0.9696	No
Maximum drawdown	-7.60%	[-19.45%, 7.72%]	0.4224	No
Calmar ratio	0.0068	[-0.5206, 0.5161]	0.8904	No

Metric BL - XGBoost	Difference	95% bootstrap CI	p-value	Significant at 5%
Turnover	157.70%	[136.31%, 183.28%]	<0.001	Yes

Source: Authors' own creation.

The paired bootstrap further shows that most full-period performance differences are not statistically significant. The annualized-return difference of 3.09 percentage points in favour of BL is not statistically significant ($p=0.5608$). Similarly, the differences in the Sharpe ratio ($p=0.9696$), maximum drawdown ($p=0.4224$), and Calmar ratio ($p=0.8904$) are not statistically significant. The statistically significant differences are annualized volatility ($p<0.001$) and turnover (<0.001). These results support a cautious interpretation of the overall comparison. XGBoost provides statistically robust reductions in volatility and turnover, while the observed differences in return and risk-adjusted performance are not statistically significant.

5.3 SHAP Explanations

In addition to the comparative analysis of both models' performance for stock portfolio optimization, we also report the global SHAP feature importances of the XGBoost model. Table 7 summarizes, for each input variable, the average absolute SHAP value and its normalized share of total importance, together with the corresponding economic category.

Table 7. SHAP Results

Metric	Economic group	Mean	Normalized importance (% of total)
atr_14	Volatility	0.013947	24.21%
sma_50	Trend	0.009150	15.89%
vol_sma_20	Volume	0.008078	14.03%
sma_20	Trend	0.007945	13.79%
ann_vol_20	Volatility	0.005301	9.20%
sma_10	Trend	0.005122	8.89%
rsi_14	Oscillator	0.002031	3.53%
macd	Oscillator	0.001819	3.16%
macd_signal	Oscillator	0.001763	3.06%
macd_hist	Oscillator	0.000807	1.40%
ret_1	Short-term return	0.000719	1.25%
vol_ratio	Volume	0.000595	1.03%
log_ret_1	Short-term return	0.000321	0.56%

Source: Authors' own creation.

The SHAP results are computed directly on the trained XGBoost ensemble using a tree-based explainer, which indicates that the explanations are faithful to the

model The most important feature, `atr_14`, accounts for about 24.21% of total importance, while the remaining percentage is spread across the other indicators. The top three and top five features explain 54.13% and 77.12%, respectively, of total importance. This indicates that the model relies on a small set of key technical drivers but is not completely based on a single variable. In this way, the explanations are faithful to the model and not entirely driven by random noise.

Furthermore, trend indicators represent the largest share of total importance (38.57%), followed by volatility (33.41%), volume (15.06%), oscillators (11.15%) and short-term returns (1.81%). In economic terms, this means that when the model predicts higher forward returns, it does so mainly in situations characterized by persistent trends, specific volatility regimes and momentum configurations, and is not primarily driven by isolated daily price moves. Thus, the XAI method produces explanations that are faithful to the model, and economically meaningful.

6. Conclusions

Our paper aimed to compare the performance of a classical portfolio-optimization model with an AI-based strategy. Both strategies were evaluated on a dataset composed of market indices, sector ETFs and stocks over a long sample of market regimes. The analysis addressed two questions: whether an AI model can consistently outperform the Black-Litterman benchmark and whether XAI tools can produce explanations that are faithful to the model and economically interpretable.

Regarding RQ1, the results do not indicate consistent superiority of either strategy. BL achieved a higher annualized return, while XGBoost produced lower volatility, a shallower maximum drawdown and substantially lower turnover. However, the differences in volatility and turnover were statistically significant, while the subperiod analysis showed that performance varies across market periods. Therefore, the results suggest a trade-off between the two approaches rather than a clear dominance of one model.

With respect to RQ2, the SHAP diagnostics for XGBoost indicate that the model's behaviour can be explained in both technically and economically meaningful ways. Feature importance is distributed across a relatively small set of indicators, dominated by trend and volatility measures, followed by volume variables and oscillators, with short-term returns playing a secondary role. Moreover, the importance is not concentrated in a single predictor, and the tree-based SHAP explainer directly reflects the fitted ensemble, suggesting that the explanations are faithful to the underlying model. For practitioners, this provides a transparent link between AI-driven forecasts and familiar technical concepts, helping to bridge the gap between black-box models and traditional portfolio intuition.

Explainability represents a key feature for investors but does not resolve all concerns associated with stock portfolio optimization. A promising direction for future research is to examine a hybrid approach combining the two models rather than replacing the BL framework entirely. One possible design would use XAI models to generate or refine the view inputs for the BL strategy.

Finally, our study has several limitations that provide directions for future research. Transaction costs are not explicitly modelled, and the analysis uses a single machine-learning model. The XGBoost model is estimated once before the out-of-sample period and is not updated through walk-forward or expanding-window retraining. The BL implementation also approximates the market-equilibrium portfolio using equal weights rather than market-capitalization weights, which may affect the estimated equilibrium prior and resulting allocations. In addition, assuming a zero risk-free rate may influence the Shape-based comparison across a sample containing different interest-rate environments.

Future studies could extend the comparison to deep learning or reinforcement-learning agents. On the explainability side, richer local explanations, user-centred evaluation with portfolio managers, and comparisons across XAI techniques could further clarify how interpretability affects trust and adoption in real investment processes. Overall, the evidence suggests that ML could complement well-established portfolio frameworks such as BL but not fully replace them yet.

References

- [1] Alzaman, C. (2024), *Deep learning in stock portfolio selection and predictions*. *Expert Systems with Applications*, 237, 121404, <https://doi.org/10.1016/j.eswa.2023.121404>.
- [2] Ashrafzadeh, M., Sadrani, M., Zolfani, S.H. (2025), *Deep learning and machine learning models for portfolio optimization: Enhancing return prediction with stock clustering*. *Results in Engineering*, 27, 106263, <https://doi.org/10.1016/j.rineng.2025.106263>.
- [3] Black, F., Litterman, R. (1992), *Global Portfolio Optimization*. *Financial Analysts Journal*, 48, 28-43, <https://doi.org/10.2469/faj.v48.n5.28>.
- [4] Černevičienė, J., Kabašinskas, A. (2024), *Explainable artificial intelligence (XAI) in finance: a systematic literature review*. *Artificial Intelligence Review*, 57, 216, <https://doi.org/10.1007/s10462-024-10854-8>.
- [5] Cerqueira, V., Torgo, L., Mozetič, I. (2020), *Evaluating time series forecasting models: an empirical study on performance estimation methods*. *Machine Learning*, 109, 1997-2028, <https://doi.org/10.1007/s10994-020-05910-7>.
- [6] Chen, T., Guestrin, C. (2016), *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paper presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, California, USA, 785-794, <https://doi.org/10.1145/2939672.2939785>.
- [7] Chincarini, L.B., Kim, D. (2013), *Uses and Misuses of the Black-Litterman Model in Portfolio Construction*. *Journal of Mathematical Finance*, 3, 153-164, <https://doi.org/10.4236/jmf.2013.31A015>.
- [8] Espiga-Fernández, F., García-Sánchez, Á., Ordieres-Meré, J. (2024), *A Systematic Approach to Portfolio Optimization: A Comparative Study of Reinforcement Learning Agents, Market Signals, and Investment Horizons*. *Algorithms*, 17, 570, <https://doi.org/10.3390/a17120570>.

- [9] Fischer, T., Krauss, C. (2018), *Deep learning with long short-term memory networks for financial market predictions*. *European Journal of Operational Research*, 270, 654-669, <https://doi.org/10.1016/j.ejor.2017.11.054>.
- [10] Freibauer, J., Grawert, S., Rieger, M.O. (2024), *The effects of trading apps on investment behavior over time*. *The European Journal of Finance*, 1-25, <https://doi.org/10.1080/1351847X.2024.2401604>.
- [11] Goswami, B., Uddin, A. (2025), *Significance of predictors: revisiting stock return predictions using explainable AI*. *Annals of Operations Research*, <https://doi.org/10.1007/s10479-025-06717-2>.
- [12] Harris, R.D.F., Stoja, E., Tan, L. (2017), *The dynamic Black–Litterman approach to asset allocation*. *European Journal of Operational Research*, 259, 1085-1096, <https://doi.org/10.1016/j.ejor.2016.11.045>.
- [13] Jang, J., Seong, N. (2023), *Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory*. *Expert Systems with Applications*, 218, 119556, <https://doi.org/10.1016/j.eswa.2023.119556>.
- [14] Jayadi, C.F., Sumarti, N. (2023), *Stock and Structured Warrant Portfolio Optimization Using Black-Litterman Model and Binomial Method*. *Mendel*, 29, 220-228, <https://doi.org/10.13164/mendel.2023.2.220>.
- [15] Kolm, P., Ritter, G. (2017), *On the Bayesian interpretation of Black–Litterman*. *European Journal of Operational Research*, 258, 564-572, <https://doi.org/10.1016/j.ejor.2016.10.027>.
- [16] Kolm, P.N., Tütüncü, R., Fabozzi, F.J. (2014), *60 Years of portfolio optimization: Practical challenges and current trends*. *European Journal of Operational Research*, 234, 356–371, <https://doi.org/10.1016/j.ejor.2013.10.060>.
- [17] Kubo, K., Nakagawa, K. (2025), *Portfolio optimization using deep learning with risk aversion utility function*. *Finance Research Letters*, 74, 106761, <https://doi.org/10.1016/j.frl.2025.106761>.
- [18] Künsch, H.R. (1989), *The Jackknife and the Bootstrap for General Stationary Observations*. *The Annals of Statistics*, 17, 1217-1241, <https://doi.org/10.1214/aos/1176347265>.
- [19] Lundberg, S.M., Lee, S.-I. (2017), *A Unified Approach to Interpreting Model Predictions*. In *Advances in Neural Information Processing Systems 30*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California, USA, 4765-4774.
- [20] Ma, Y., Han, R., Wang, W. (2021), *Portfolio optimization with return prediction using deep learning and machine learning*. *Expert Systems with Applications*, 165, 113973, <https://doi.org/10.1016/j.eswa.2020.113973>.
- [21] Palczewski, A., Palczewski, J. (2019), *Black–Litterman model for continuous distributions*. *European Journal of Operational Research*, 273, 708-720, <https://doi.org/10.1016/j.ejor.2018.08.013>.
- [22] Sahamkhadam, M., Stephan, A., Östermark, R. (2022), *Copula-based Black–Litterman portfolio optimization*. *European Journal of Operational Research*, 297, 1055-1070, <https://doi.org/10.1016/j.ejor.2021.06.015>.

- [23] Satchell, S., Scowcroft, A. (2000), *A demystification of the Black–Litterman model: Managing quantitative and traditional portfolio construction*. *Journal of Asset Management*, 1, 138-150, <https://doi.org/10.1057/palgrave.jam.2240011>.
- [24] Simos, T.E., Mourtas, S.D., Katsikis, V.N. (2021), *Time-varying Black–Litterman portfolio optimization using a bio-inspired approach and neuronets*. *Applied Soft Computing*, 112, 107767, <https://doi.org/10.1016/j.asoc.2021.107767>.
- [25] Soleymani, F., Paquet, E. (2021), *Deep graph convolutional reinforcement learning for financial portfolio management – DeepPocket*. *Expert Systems with Applications*, 182, 115127, <https://doi.org/10.1016/j.eswa.2021.115127>.
- [26] Sun, Y., Wu, Y., De, G. (2023), *A Novel Black-Litterman Model with Time-Varying Covariance for Optimal Asset Allocation of Pension Funds*. *Mathematics*, 11, 1476, <https://doi.org/10.3390/math11061476>.
- [27] Sutiene, K., Schwendner, P., Sipos, C., Lorenzo, L., Mirchev, M., Lameski, P., Kabasinskas, A., Tidjani, C., Ozturkkal, B., Cerneviene, J. (2024), *Enhancing portfolio management using artificial intelligence: literature review*. *Frontiers in Artificial Intelligence*, 7, 1371502, <https://doi.org/10.3389/frai.2024.1371502>.
- [28] Weber, P., Carl, K.V., Hinz, O. (2024), *Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature*. *Management Review Quarterly*, 74, 867-907, <https://doi.org/10.1007/s11301-023-00320-0>.
- [29] Yang, S. (2023), *Deep reinforcement learning for portfolio management*. *Knowledge-Based Systems*, 278, 110905, <https://doi.org/10.1016/j.knosys.2023.110905>.
- [30] Yeo, W.J., Van Der Heever, W., Mao, R., Cambria, E., Satapathy, R., Mengaldo, G. (2025), *A comprehensive review on financial explainable AI*. *Artificial Intelligence Review*, 58, 189, <https://doi.org/10.1007/s10462-024-11077-7>.

Appendix

Table A1

Period	Model	Return	Volatility	Sharpe	MDD	Calmar	Turnover
2018-2019	BL	11.30%	15.45%	0.731	-17.18%	0.658	235.78%
	XGBoost	9.97%	14.82%	0.673	-17.48%	0.571	57.33%
2020-2021	BL	36.69%	28.40%	1.292	-30.45%	1.205	206.50%
	XGBoost	15.26%	18.02%	0.847	-26.87%	0.568	52.22%
2022-2025	BL	3.54%	17.90%	0.198	-34.43%	0.103	195.74%
	XGBoost	7.57%	15.37%	0.492	-23.94%	0.316	46.48%