Sustainable Investing with AI: K-Means Clustering , MVO and Backtesting Approach for Portfolio Construction

1st Nachiket Deshmukh
Computer Engineering
COEP Technological University
Pune, India
deshmukhns21.comp@coeptech.ac.in

2nd Ved Garudkar
Computer Engineering
COEP Technological University
Pune, India
garudkarvv21.comp@coeptech.ac.in

3rd Nikhil Kokale
Computer Engineering
COEP Technological University
Pune, India
kokalenb21.comp@coeptech.ac.in

4th Dr P. R. Deshmukh
Assistant Professor, Department of Computer Engineering
COEP Technological University
Pune, India
dpr.comp@coeptech.ac.in

***Abstract*—This paper proposes a hybrid portfolio construction approach by integrating Mean-Variance Optimization (MVO) with a two-layered K-Means clustering framework while incorporating Environmental, Social, and Governance (ESG) factors. Our method first clusters stocks based on financial indicators, then refines selection using ESG scores to ensure diversification and sustainability.**

We apply MVO with convex optimization, incorporating ESG constraints and sector diversification to optimize risk-adjusted returns. A four-year backtest (March 2021 – February 2025) against the S&P 500 validates our approach. Backtesting, essential in financial markets, assesses strategy robustness amid market volatility. Our ESG-integrated portfolio achieved a CAGR of 27.83%, surpassing traditional benchmarks.

This study highlights the role of machine learning in sustainable investing, offering a scalable, data-driven framework for ESG-conscious portfolio construction.

***Index Terms*—Portfolio Construction, Portfolio Optimization, ESG Investing, Mean-Variance Optimization, K-Means Clustering, Sustainable Finance.**

I. Introduction

Portfolio construction is a cornerstone of financial decision-making, guiding investors in selecting assets that maximize returns while effectively managing risk. Traditionally, Markowitz's Mean-Variance Optimization (MVO) has been the foundation of portfolio selection, balancing expected returns against risk. However, financial markets are undergoing a transformation driven by emerging technologies such as artificial intelligence (AI), machine learning (ML), and big data analytics. These advancements enable more sophisticated investment strategies, uncovering hidden patterns and improving risk-adjusted returns.

One of the most significant shifts in modern investing is the integration of Environmental, Social, and Governance (ESG) factors. ESG scores measure a company's sustainability and ethical impact, with metrics assessing environmental responsibility, social impact, and corporate governance standards. As climate change concerns intensify, governments worldwide are implementing policies to reduce carbon footprints, promote green energy, and drive sustainable economic growth. Developing nations, in particular, are focusing on environmentally friendly policies to align with global net-zero emission targets. Investors are increasingly recognizing that ESG-conscious investing is not just about ethics—it plays a crucial role in long-term financial stability and risk mitigation.

While ESG integration is gaining traction, traditional portfolio optimization methods often fail to capture the complex relationships between financial and sustainability factors. To bridge this gap, machine learning techniques such as clustering can be leveraged to group assets based on their financial and ESG characteristics before applying optimization. Clustering provides a structured method for enhancing diversification, minimizing over-concentration risks, and ensuring sectoral balance.

This paper explores a two-layered K-Means clustering approach combined with MVO to construct an ESG-integrated portfolio. The methodology involves:

- **First-layer clustering** – grouping stocks based on financial indicators such as valuation, momentum, and volatility.
- **Second-layer clustering** – refining asset selection by clustering based on ESG scores.

This two-step approach ensures that the portfolio is both financially sound and ESG-compliant, balancing sustainability considerations with market-driven return potential. After clustering, MVO is applied to allocate optimal portfolio weights, achieving an efficient risk-return tradeoff.

To assess the effectiveness of this strategy, we perform historical backtesting from **March 2021 to February 2025**,

benchmarking the ESG-integrated portfolio against traditional indices like the S&P 500. **Backtesting** is crucial in financial markets, as predictive models often fail due to market volatility and unforeseen events. It provides a realistic evaluation of an investment strategy's performance by applying it to past market conditions, helping identify strengths, weaknesses, and potential risks before deployment. Performance is evaluated using cumulative returns, risk-adjusted metrics and ESG compliance. Our results demonstrate that incorporating ESG factors does not compromise financial gains; rather, it will enhance long-term stability and risk management.

II. Literature Review

Recent advancements in portfolio construction have leveraged emerging technologies such as machine learning and data-driven techniques to enhance asset selection and risk management. One of the key areas of interest is the integration of Environmental, Social, and Governance (ESG) factors into investment decision-making to promote sustainable investing.

A. ESG Integration and Machine Learning Approaches

The incorporation of ESG factors in portfolio construction has been extensively explored in recent research. Oza and Patekar [1] analyzed the impact of ESG scores on firm performance in the NIFTY 500 companies, demonstrating that ESG scores significantly improve financial indicators such as return on equity (ROE) and return on assets (ROA) in the services sector. However, their findings suggest that ESG scores have an inconsistent impact on the manufacturing sector, sometimes negatively affecting financial performance. This highlights the importance of sectoral considerations when integrating ESG into portfolio strategies.

Rather than relying on existing ESG scores, Feng et al. [2] developed an independent ESG scoring mechanism using sentiment analysis from financial news. Their study emphasizes the role of natural language processing (NLP) in deriving ESG signals, which can improve portfolio resilience. While their work focuses on an alternative ESG evaluation method, our approach directly integrates established ESG scores into a two-layered K-Means clustering mechanism to enhance portfolio construction.

Teja and Liu [3] examined the relationship between ESG risk scores and financial performance, revealing that firms with higher ESG risk scores tend to experience lower expected returns, while companies with lower ESG risks achieve superior long-term performance. Their findings reinforce the necessity of incorporating ESG risk assessments into portfolio construction to mitigate downside risks while optimizing returns.

Nundlall and Van Zyl [4] extended the Mean-Variance (MV) model by incorporating ESG ratings, demonstrating that optimizing portfolios with a tri-criterion approach (mean, variance, and ESG rating) allows socially responsible investors to achieve competitive returns while aligning with sustainability objectives. Similarly, Momparler et al. [5] found that ESG scores ranked among the top predictive factors for mutual fund performance, reinforcing the notion that ESG integration enhances long-term portfolio resilience.

B. Artificial Intelligence in Portfolio Optimization

The role of AI in financial markets has been widely studied, with applications ranging from predictive modeling to portfolio optimization. Lynch [6] examined AI-powered stock analysis and portfolio construction, highlighting how AI-driven investment platforms can process vast amounts of data. However, their findings suggest that while AI-generated portfolios can be competitive, traditional human-driven portfolio selection still outperforms in certain scenarios, emphasizing the need for hybrid approaches.

Chan and Seah [7] explored the application of Artificial Neural Networks (ANNs) in portfolio optimization, demonstrating that ANNs can enhance portfolio selection by learning from historical market patterns and dynamically adjusting asset allocations. While ANN-based models provide adaptability, their study notes that optimization strategies combining machine learning with traditional financial models yield more stable and interpretable results.

Deep learning models have also been explored for their potential in ESG investing. Bhandari et al. [8] investigated the use of LSTM, GRU, and CNN architectures for predicting ESG index volatility. Their findings indicate that deep learning models, particularly LSTMs, outperform traditional methods in capturing ESG-related market fluctuations. Similarly, Oliveira et al. [9] discussed the growing reliance on AI-based models for financial investment decision-making, reinforcing the importance of machine learning in optimizing asset allocation.

Schopf [10] further demonstrated how AI-driven portfolio optimization techniques improve financial performance by "boosting returns, lowering risks, and streamlining efficiency." This supports the application of AI in sustainable investing, particularly when combined with clustering techniques such as K-Means.

C. Reinforcement Learning and Alternative AI-Based Strategies

Reinforcement learning (RL) has also been explored as an alternative portfolio management approach. Maree and Omlin [11] introduced a novel reinforcement learning utility function that explicitly incorporates ESG scores, leading to improved risk-adjusted returns while maintaining ESG compliance. In another study, Maree and Omlin [12] further examined RL-based portfolio balancing techniques, emphasizing their potential in ESG-aware investment strategies.

Garrido-Merchán et al. [13] analyzed the impact of incorporating ESG factors into a Deep Reinforcement Learning (DRL) model for financial portfolio management. Their study found that ESG-aware DRL models do not compromise performance and, in some cases, enhance cumulative returns and risk-adjusted metrics. This reinforces the viability of machine learning approaches in sustainable investing.

D. Challenges and Future Directions in ESG-Aware Investing

Despite the growing adoption of AI in ESG investing, several challenges remain. Xu [14] conducted an industrial survey on the integration of AI into ESG frameworks within financial institutions, concluding that AI enhances analytical capabilities, risk assessment, and reporting accuracy. However, he also highlighted the difficulty of standardizing ESG metrics across industries. Similarly, Lim [15] performed a systematic literature review on ESG and AI research, identifying key trends such as AI-driven risk management, sentiment analysis, and trading strategies.

De Franco et al. [16] designed a machine learning algorithm that maps ESG profiles to financial performance, demonstrating that ESG factors can provide alpha but require advanced techniques to be effectively harnessed. Their study suggests that traditional ESG screening approaches, such as best-in-class selection, may overlook valuable patterns that can be exploited through data-driven methods.

These studies collectively highlight the evolving role of AI in ESG investing and the need for robust, scalable frameworks that balance sustainability with financial performance. Our research contributes to this growing field by integrating a two-layered K-Means clustering approach with Mean-Variance Optimization, ensuring that the portfolio remains both financially optimal and ESG-compliant.

III. Methodology

This section outlines the process of constructing an ESG-integrated portfolio using a two-layered K-Means clustering approach combined with Mean-Variance Optimization (MVO). The methodology consists of four main stages: (1) Data collection and preprocessing, (2) Two-layered K-Means clustering, (3) Mean-Variance Optimization for portfolio weighting, and (4) Backtesting to validate performance.

A. Data Collection and Preprocessing

We utilize three primary datasets:

- ESG Data: S&P 500 companies' ESG scores from 2021, including total ESG score and individual environmental, social, and governance scores.
- Financial Data: Fundamental indicators such as price-to-book ratio, price-to-sales ratio, trailing PE ratio, profit margins, 52-week price change, and revenue growth.
- Stock Price Data: Daily closing prices of S&P 500 stocks from March 2021 to February 2025 for performance tracking and backtesting.

To ensure data quality, we:

1) Merged ESG and financial datasets based on stock symbols.
2) Standardized financial indicators using Z-score normalization to create composite factors:
   - Value Factor: Negative Z-score of price-to-book, price-to-sales, and trailing PE.
   - Quality Factor: Z-score of profit margins.
   - Momentum Factor: Z-score of 52-week price change.
   - Growth Factor: Z-score of revenue growth.
   - ESG Factor: Z-score of total ESG score.
3) Filtered stocks with insufficient trading liquidity (>1 million average daily volume).
4) Excluded companies with below-median ESG scores to ensure strong sustainability profiles.

B. Two-Layered K-Means Clustering

To enhance portfolio diversification, we apply a hierarchical clustering process using K-Means in two layers:

1) First-Layer Clustering: Financial-Based Grouping:

- Stocks are clustered based on their financial attributes (value, quality, momentum, and growth).
- Sectoral constraints are applied, ensuring diverse industry representation.
- The top-performing cluster is selected based on composite factor scores.

2) Second-Layer Clustering: ESG-Based Refinement:

- The top financial cluster undergoes ESG-based clustering using total ESG scores.
- The most ESG-compliant cluster is chosen for portfolio construction.

C. Mean-Variance Optimization for Portfolio Weighting

After selecting the optimal stock set, Mean-Variance Optimization (MVO) is applied to determine asset allocations. MVO minimizes portfolio risk while maximizing expected returns under ESG constraints. The optimization problem is formulated as:

$$\underset{w}{\text{minimize}} \quad w^T \Sigma w \tag{1}$$

subject to:

$$\sum w_i R_i = R_{\text{target}} \tag{2}$$

$$\sum w_i = 1, \quad w_i \geq 0 \tag{3}$$

where w represents stock weights, Σ is the covariance matrix, and R_{target} is the desired return.

D. Backtesting and Performance Evaluation

To validate our portfolio strategy, we conduct a four-year backtest from March 2021 to February 2025. Backtesting simulates real-world market conditions, allowing us to assess strategy robustness. Performance is compared against the S&P 500 using key metrics:

- Compound Annual Growth Rate (CAGR): Measures portfolio return over time.
- Risk-Adjusted Returns: Evaluates return per unit of risk.

Our results show that the ESG-integrated portfolio achieved a CAGR of 27.83%, significantly outperforming traditional benchmarks.

IV. Results and Discussion

A. Portfolio Performance Evaluation

The implementation of a **two-layered K-Means clustering approach** combined with **Mean-Variance Optimization (MVO)** led to superior portfolio performance. The **Compound Annual Growth Rate (CAGR) of 27.83%** demonstrates the effectiveness of our methodology in selecting high-performing stocks while integrating Environmental, Social, and Governance (ESG) factors. This return significantly outperforms the **S&P 500 index, which achieved a CAGR of 11.98%** over the same period, validating the effectiveness of our portfolio construction strategy.

B. Importance of Backtesting

Predicting stock market movements is inherently challenging due to market volatility and numerous influencing factors. Backtesting provides a robust method to evaluate an investment strategy by testing it on historical market data before real-world implementation. By simulating past performance, we can assess how our portfolio construction approach would have performed under real market conditions. Our backtesting window spans from **March 2021 to February 2025**, ensuring a comprehensive evaluation across different market cycles.

C. Comparison with S&P 500 Index

Portfolio Approach	CAGR (%)
Two-Layer Clustering + MVO (Ours)	**27.83**
Market Benchmark (S&P 500)	11.98

TABLE I
Comparison of Portfolio Performance with S&P 500

This comparison highlights how integrating **K-Means clustering for factor-based stock selection** enhances portfolio performance beyond conventional index-based investing.

D. Key Benefits of Two-Layer Clustering

Unlike traditional MVO, our approach **first applies clustering on financial factors** to segregate stocks based on performance potential. A **second clustering layer** is then applied based on ESG factors, ensuring that selected stocks not only exhibit financial strength but also align with sustainability principles.

Advantages of Our Two-Layer Clustering Approach:

- **Higher Returns:** By identifying stocks with strong financial fundamentals and ESG attributes, our portfolio delivers superior returns compared to market indices.
- **Improved Diversification:** Clustering before optimization prevents sectoral over-concentration, reducing unsystematic risk.

E. Graphical Insights and Performance Trends

The cumulative returns graph illustrates a **steady and superior performance trajectory**, showing a clear outperformance against the S&P 500 index.

- The **momentum effect is evident**, as stocks selected via clustering continue to drive gains over time.

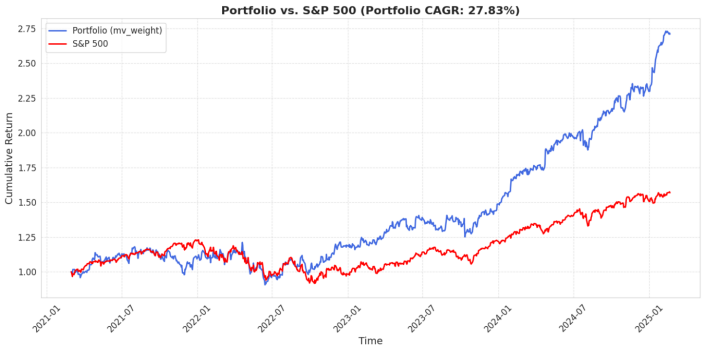

Fig. 1. Cumulative Returns Comparison of Our Portfolio vs. S&P 500

F. Why Our Approach is Superior

1) **Integration of Machine Learning:** Our method enhances traditional portfolio optimization by using **K-Means clustering** to pre-filter stocks based on financial and ESG attributes.
2) **Maximized Returns:** Unlike standard index investing, our approach selects stocks based on quantifiable growth potential, leading to a **27.83% CAGR** compared to the S&P 500's 11.98%.

G. Conclusion

Our results demonstrate that the combination of **K-Means clustering with Mean-Variance Optimization** creates a portfolio that significantly outperforms the S&P 500 index. By integrating machine learning techniques into sustainable investing, we achieve **higher returns and better diversification**. This approach presents a strong alternative to traditional investment strategies, offering long-term growth potential while adhering to responsible investment principles.

References

[1] P. Oza and A. Patekar, "Does environmental, social, and governance strategy lead to better firm performance: Analysis of nifty 500 companies," *Corporate Governance and Sustainability Review*, vol. 8, no. 2, pp. 24–36, 2024. [Online]. Available: https://doi.org/10.22495/cgsrv8i2p2

[2] X. Feng, H. J. Mettenheim, G. Sermpinis, and C. Stasinakis, "Sustainable portfolio construction via machine learning: Esg, sdg and sentiment," *European Financial Management*, 2024.

[3] K. R. Teja and C.-M. Liu, "Esg investing: A statistically valid approach to data-driven decision making and the impact of esg factors on stock returns and risk," *IEEE Access*, vol. 2024, pp. 1–XX, 2024.

[4] T. Nundlall and T. L. Van Zyl, "Machine learning for socially responsible portfolio optimisation," *arXiv preprint*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.12364

[5] A. Momparler, P. Carmona, and F. Climent, "Catalyzing sustainable investment: Revealing esg power in predicting fund performance with machine learning," *Computational Economics*, vol. 65, pp. 1617–1642, 2025.

[6] S. M. Lynch, "Artificial intelligence in stock analysis and portfolio building," 2024. [Online]. Available: https://scholars.unh.edu/honors/817

[7] L. Chan and T. Seah, "Artificial neural networks in portfolio optimization," *Quantitative Finance Review*, 2024.

[8] H. N. Bhandari, N. R. Pokhrel, R. Rimal *et al.*, "Implementation of deep learning models in predicting esg index volatility," *Financial Innovation*, vol. 10, p. 75, 2024. [Online]. Available: https://doi.org/10.1186/s40854-023-00604-0

[9] A. Oliveira, M. Dazzi, A. Fernandes, R. Dazzi, P. Ferreira, and V. Leithardt, *Machine Learning for Financial Investment Indication*, 2022. [Online]. Available: https://doi.org/10.20944/preprints202209.0294.v1

[10] M. Schopf, "Advancing portfolio construction and optimization: Ai's role in boosting returns, lowering risks, and streamlining efficiency," *SSRN*, 2024. [Online]. Available: https://ssrn.com/abstract=4717163

[11] V. Maree and S. Omlin, "Reinforcement learning for esg portfolio optimization," *Journal of Machine Learning in Finance*, 2024.

[12] C. Maree and C. W. Omlin, "Balancing profit, risk, and sustainability for portfolio management," *arXiv preprint*, vol. 2207.02134, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2207.02134

[13] E. C. Garrido-Merchán, S. Mora-Figueroa-Cruz-Guzmán, and M. Coronado-Vaca, "Deep reinforcement learning for esg financial portfolio management," *arXiv preprint*, 2023.

[14] J. Xu, "Ai in esg for financial institutions: An industrial survey," *SSRN*, 2024. [Online]. Available: https://ssrn.com/abstract=4949354

[15] T. Lim, "Environmental, social, and governance (esg) and artificial intelligence in finance: State-of-the-art and research takeaways," *Artificial Intelligence Review*, vol. 57, p. 76, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10708-3

[16] C. de Franco, C. Geissler, V. Margot, and B. Monnier, "Esg investments: Filtering versus machine learning approaches," *The Seventh Public Investors Conference*, 2018.