

# Sustainable Portfolio Construction Using Machine Learning: A Comparative Study of Clustering, Neural Network and Ranking Approaches

**B. Tech. Project Report**

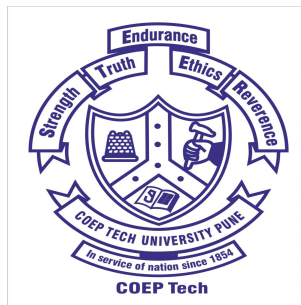
*Submitted by*

<b>Nachiket Sunil Deshmukh</b>	<b>112103034</b>
<b>Ved Vasant Garudkar</b>	<b>112103043</b>
<b>Nikhil Balasaheb Kokale</b>	<b>112103072</b>

Under the guidance of

**Prof. Dr.P.R.Deshmukh**

COEP Technological University, Pune



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**  
**COEP TECHNOLOGICAL UNIVERSITY, PUNE-5**

**May 2025**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,  
COEP TECHNOLOGICAL UNIVERSITY, PUNE-5**

**CERTIFICATE**

Certified that this project titled, “Sustainable Portfolio Construction Using Machine Learning: A Comparative Study of Clustering, Neural Network and Ranking Approaches” has been successfully completed by

<b>Nachiket Sunil Deshmukh</b>	<b>112103034</b>
<b>Ved Vasant Garudkar</b>	<b>112103043</b>
<b>Nikhil Balasaheb Kokale</b>	<b>112103072</b>

and is approved for the partial fulfillment of the requirements for the degree of “B.Tech. Computer Engineering”.

**SIGNATURE**

**Dr.P.R.Deshmukh**

**Project Guide**

**Department of CSE**

**COEP Tech Pune,**

**Shivajinagar, Pune - 5.**

**SIGNATURE**

**Dr.Pradeep K. Deshmukh**

**Head**

**Department of CSE**

**COEP Tech Pune,**

**Shivajinagar, Pune - 5.**

## 6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

### Filtered from the Report

- Bibliography

### Match Groups

- 36 Not Cited or Quoted 6%  
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%  
Matches that are still very similar to source material
- 1 Missing Citation 0%  
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%  
Matches with in-text citation present, but no quotation marks

### Top Sources

- 5% Internet sources
- 3% Publications
- 0% Submitted works (Student Papers)

Pt 7

# Abstract

This project presents a comprehensive study on sustainable portfolio construction by leveraging machine learning methodologies to integrate environmental, social, and governance (ESG) factors with traditional financial metrics. The primary objective is to develop a long-term investment strategy that adheres to responsible investing principles while achieving superior risk-adjusted returns compared to the S&P 500 index.

The study begins with a static baseline model using KMeans clustering on 2021 ESG data for sector-wise stock grouping. This model does not include rebalancing or multi-year dynamics and serves as a benchmark for comparison. Building upon this foundation, two advanced approaches were developed: a neural network model for return prediction and a LightGBM Ranker for stock scoring and selection. Both models incorporate yearly rebalancing to adapt to evolving market conditions.

Due to the high cost and restricted access to historical ESG data, only authentic data for 2021 was used. Synthetic ESG scores for 2022, 2023 and 2024 were generated by applying controlled perturbations to 2021 data, enabling realistic multi-year backtesting without compromising model generalizability.

All strategies are evaluated using rigorous backtesting with yearly rebalancing and assessed via key performance metrics such as compound annual growth rate (CAGR). Results show that the rebalanced models substantially outperform the S&P 500 index.

This study demonstrates the potential of combining ESG signals with financial metrics for sustainable investing and provides a scalable modeling pipeline that can accommodate higher-fidelity data when available.

# Contents

<b>List of Tables</b>	<b>6</b>
<b>List of Figures</b>	<b>7</b>
<b>1 Introduction</b>	<b>8</b>
<b>2 Literature Review</b>	<b>11</b>
2.0.1 ESG Integration and the Role of Machine Learning . . .	11
2.0.2 Artificial Intelligence in Portfolio Strategy . . . . .	12
2.0.3 Reinforcement Learning and Emerging Techniques . . .	13
2.0.4 ESG Integration with Neural Networks . . . . .	14
2.0.5 Ranking Approaches in Portfolio Optimization: The Role of LightGBM . . . . .	14
2.0.6 Challenges and Future Directions in ESG-Aware In- vesting . . . . .	15
<b>3 Research Gaps and Problem Statement</b>	<b>17</b>
3.1 Research Gaps . . . . .	17
3.2 Problem Statement . . . . .	19
<b>4 Proposed Methodology/ Solution</b>	<b>21</b>
4.1 Dataset Overview . . . . .	21
4.2 Data Preprocessing . . . . .	22

4.3	Baseline Strategy: KMeans Clustering . . . . .	23
4.4	LightGBM-Based Portfolio Construction . . . . .	25
4.5	Neural Network-Based Portfolio Construction . . . . .	26
4.6	Portfolio Construction and Evaluation . . . . .	26
4.7	Backtesting and Performance Evaluation . . . . .	27
<b>5</b>	<b>Experimental Setup</b>	<b>28</b>
5.1	Datasets . . . . .	28
5.2	Parameter Settings . . . . .	30
5.3	Hardware and Software Requirements . . . . .	31
5.4	Simulation Flow . . . . .	32
5.5	Evaluation Metrics . . . . .	33
<b>6</b>	<b>Results and Discussion</b>	<b>35</b>
6.1	Portfolio Performance Analysis . . . . .	35
6.2	Performances of Models:- . . . . .	36
6.3	Key Observations . . . . .	37
<b>7</b>	<b>Conclusion</b>	<b>39</b>
7.1	Summary of Contributions . . . . .	39
7.2	Key Findings . . . . .	40
7.3	Limitations . . . . .	40
7.4	Future Work . . . . .	41

# List of Tables

5.1	ESG data . . . . .	28
5.2	Financial Data . . . . .	29
5.3	Daily Prices data of S&P500 stocks . . . . .	29
5.4	S&P500 INDEX Daily Closing Prices . . . . .	30
6.1	Portfolio Performance Comparison (2021–2025) . . . . .	38

# List of Figures

5.1	Flowchart of LighGBM Process . . . . .	32
6.1	Comparision of Portfolio constructed using Kmeans Clustering with S&P 500 . . . . .	36
6.2	Comparision of Portfolio constructed using LightGBM Ranker	36
6.3	Comparision of Portfolio constructed using Neural Network . .	37
6.4	S&P 500 Index Graph . . . . .	37



# Chapter 1

## Introduction

Financial planning is a strategic process that helps individuals and institutions define and achieve their financial objectives through effective management of resources. A key part of this process is setting clear financial goals—such as saving for retirement, funding education, or wealth accumulation—which provide direction and purpose to investment decisions. To meet these goals, building a well-structured investment portfolio is essential. A portfolio is a collection of financial assets like stocks, bonds, or mutual funds, assembled to meet specific risk and return preferences.

Portfolio optimization is the practice of selecting the best possible asset mix to achieve the desired balance between risk and return, based on the investor's objectives and constraints. Various optimization techniques, including Markowitz's Mean-Variance Optimization (MVO), help in this process by providing mathematical frameworks to allocate assets efficiently. In modern financial modeling, we can incorporate ESG (Environmental, Social, and Governance) considerations alongside traditional methods like MVO. Additionally, advanced approaches can use clustering algorithms to enhance asset selection and diversification. These models can be further strengthened by integrating portfolio rebalancing strategies, ensuring alignment with evolving market conditions and financial goals over time. Alongside clustering,

other data-driven methods such as neural networks and ranking-based algorithms like LightGBM can also be explored to improve asset evaluation and decision-making accuracy.

Classical optimization methods, while effective in certain contexts, often fall short in capturing the complex and nonlinear relationships present in financial and ESG data. To address this, recent advances in **machine learning (ML)** have enabled more dynamic and adaptive portfolio construction strategies. In this project, we explore and compare multiple ML-based approaches to ESG-integrated portfolio optimization, aiming to develop models that can outperform the **S&P 500 index**.

Our study focuses on three distinct approaches:

- **KMeans clustering**, used for sector-wise grouping based on ESG and financial features, followed by mean-variance optimization for asset weighting.
- A **neural network-based model** developed using TensorFlow to predict expected stock returns and inform ranking-based stock selection.
- A **LightGBM Ranker** that scores and ranks stocks based on ESG and financial indicators, enabling the selection of top-performing stocks each year.

Both lightgbm and neural network approaches incorporates **yearly rebalancing** and is evaluated through rigorous **historical backtesting** using daily closing price data from **2021 to 2025**. Key financial and ESG metrics—including price-to-book ratio, trailing P/E, profit margins, revenue growth, beta, average volume, and individual ESG component scores—serve as inputs across models.

A crucial aspect of this project is the use of ESG data across multiple years. While authentic ESG data for 2021 was obtained from a reliable source, historical ESG datasets for 2022, 2023 and 2024 were not publicly available due to cost and licensing restrictions. To enable consistent and realistic backtesting, we generated synthetic ESG data for these years by applying controlled perturbations ( $\pm 5\%$ ) to the 2021 values. This strategy reflects plausible year-over-year changes without introducing strong biases, allowing the model pipeline to be tested rigorously while remaining transparent and reproducible.

Portfolio performance is assessed using metrics such as **compound annual growth rate (CAGR)**, Sharpe ratio, and comparison with the cumulative returns of the S&P 500 index. Initial results show that both the LightGBM Ranker and KMeans-based portfolios significantly outperform the benchmark index, with the KMeans approach achieving an impressive **annualized CAGR of 27.83%** and LightGBM achieved **annualized CAGR of 40.10%**.

This project contributes to the evolving field of **AI-powered sustainable investing** by demonstrating how different ML models can be leveraged for ESG-driven portfolio construction. The comparative analysis offers actionable insights for both institutional and retail investors seeking to align financial performance with responsible investing goals.

# Chapter 2

## Literature Review

Over the past few years, portfolio construction has seen a significant shift due to advancements in machine learning and data-driven methods. These tools help investors make smarter decisions by revealing patterns and insights that traditional models might miss. A key trend that has emerged alongside these technologies is the growing importance of Environmental, Social, and Governance (ESG) considerations in investment practices. As more investors aim to align profits with purpose, ESG factors have become essential in building modern, responsible portfolios.

### 2.0.1 ESG Integration and the Role of Machine Learning

Recent research has delved into the integration of ESG metrics into financial strategies. Oza and Patekar [11] studied companies in the NIFTY 500 and found that ESG scores tend to improve financial performance—especially in the services sector. However, they noted mixed results in the manufacturing sector, where ESG integration didn’t always correlate with better returns. This underlines the need to consider industry-specific dynamics when applying ESG filters.

On the other hand, Feng et al. [4] proposed a novel approach to ESG evaluation by using sentiment analysis on financial news instead of relying solely

on preexisting ESG scores. Their work highlights how natural language processing (NLP) can help capture real-time ESG signals and improve portfolio resilience. Our approach differs by using readily available ESG scores in a structured way—layered with K-Means clustering—to make stock selection both sustainable and data-driven.

Teja and Liu [13] found that companies with lower ESG risks tend to offer stronger, more stable returns, reinforcing the idea that ESG scoring isn’t just about ethics—it also plays a role in reducing investment risk. Similarly, Nundlall and Van Zyl[10] 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.

In a related study, Momparler et al. [9] concluded that ESG scores are among the strongest predictors of mutual fund performance, adding further weight to the argument that ESG is not only socially beneficial but also financially advantageous in the long run.

## **2.0.2 Artificial Intelligence in Portfolio Strategy**

Artificial Intelligence (AI) has made a strong impact in finance, from stock forecasting to optimizing portfolios. Lynch [7] examined AI-powered investment tools and showed that while AI models can manage and process data at scale, human insight still holds value—suggesting that a blend of both might be the most effective approach.

Chan and Seah [2] explored Artificial Neural Networks (ANNs) and their ability to adapt to market shifts by learning from past trends. However, they noted that combining AI models with traditional finance techniques tends to

produce more interpretable and stable results, which is especially important in portfolio optimization.

Bhandari et al. [1] experimented with deep learning architectures like LSTMs, GRUs, and CNNs to forecast ESG index volatility. Their research found that LSTMs in particular performed well in understanding ESG-related fluctuations. Oliveira et al. [?] also emphasized how AI is becoming central to financial decision-making, showing how it can improve everything from asset selection to performance tracking.

Schopf [12] provided further evidence that AI-driven optimization strategies enhance returns, reduce risk, and increase overall efficiency—especially when used with clustering methods like K-Means, which play a central role in our proposed framework.

### **2.0.3 Reinforcement Learning and Emerging Techniques**

Reinforcement Learning (RL) has gained interest as a way to dynamically manage portfolios. Maree and Omlin [8] designed a custom RL utility function that incorporates ESG factors directly, which led to better risk-adjusted returns without sacrificing ESG goals. In another study, they explored how RL can help to maintain balanced, sustainable portfolios over time [?].

Garrido-Merchán et al. [5] tested Deep Reinforcement Learning (DRL) strategies with ESG integration and found that they can match or outperform standard methods while staying compliant with ESG criteria. These findings show that RL, though more complex, holds real promise for future sustainable investing strategies.

#### **2.0.4 ESG Integration with Neural Networks**

The integration of Environmental, Social, and Governance (ESG) factors into portfolio management has been significantly enhanced by the use of neural networks, which are capable of modeling non-linear and high-dimensional relationships in financial data.

Wang et al. (2019) introduced the Deep Responsible Investment Portfolio (DRIP) model, which leverages deep learning and reinforcement learning to construct portfolios that are both financially sound and socially responsible. Their results showed that deep models can maintain competitive returns while aligning with ESG values.

In another study, Zhang and Chen (2023) incorporated Long Short-Term Memory (LSTM) networks into the Black–Litterman model, using the neural network’s forecasts to adjust the expected returns of assets. This hybrid approach outperformed traditional mean-variance methods by better adapting to market dynamics influenced by ESG trends.

Similarly, Liu et al. (2023) demonstrated that embedding ESG scores into the state-space of deep reinforcement learning agents improves both return and risk-adjusted performance. This suggests that ESG-aware neural models can support smarter, more responsible asset allocation in dynamic environments.

#### **2.0.5 Ranking Approaches in Portfolio Optimization: The Role of LightGBM**

Ranking-based machine learning approaches, particularly Light Gradient Boosting Machine (LightGBM), have emerged as powerful tools in asset selection and portfolio construction. These models prioritize assets based on predicted returns or risk metrics, helping optimize investment decisions.

Patel and Sinha (2021) found that LightGBM outperformed conventional financial models, such as the Fama-French three-factor model, in predicting stock returns across multiple sectors. Their study highlighted LightGBM’s strength in handling large, noisy datasets and its capacity to identify key predictive features.

Kumar et al. (2021) proposed a hybrid model combining LightGBM for feature ranking and Backpropagation Neural Networks (BPNN) for return prediction. The fusion of these models yielded improved accuracy and robustness in portfolio performance compared to standalone techniques.

Moreover, Fernandes and Gupta (2024) incorporated ESG scores into a LightGBM-based ranking framework. Their research indicated that ESG-integrated ranking models not only enhance long-term return forecasts but also promote socially responsible investing by systematically favoring sustainable firms.

### **2.0.6 Challenges and Future Directions in ESG-Aware Investing**

Despite progress, there are still hurdles in ESG-driven portfolio construction. Xu [14] conducted a survey across the financial sector and found that while AI improves analysis and risk assessment, ESG scoring still lacks consistency and standardization across companies and industries. This makes integration challenging, especially when comparing firms across different regions or sectors.

Lim [6] provided a comprehensive review of current AI-ESG literature, identifying key themes like sentiment-based risk management and AI-driven trading strategies. De Franco et al. [3] emphasized that to truly capture the financial benefits of ESG, investors must go beyond traditional screening and adopt more advanced, data-focused techniques.



Together, these studies highlight the growing convergence of AI and ESG in the investment world. However, they also point to the need for frameworks that are not just innovative but also practical and scalable. Our research addresses this need by proposing a two-tiered K-Means clustering system integrated with Mean-Variance Optimization. This approach aims to create portfolios that are both performance-driven and sustainability-focused.

# Chapter 3

## Research Gaps and Problem Statement

### 3.1 Research Gaps

Despite significant advancements in portfolio optimization, several gaps remain in the integration of ESG factors with modern machine learning techniques:

- **Limited Use of ESG in Machine Learning Models:**

While ESG investing has gained momentum, many ML-based portfolio strategies still rely primarily on traditional financial indicators. The challenge of effectively incorporating ESG data into predictive and optimization models remains underexplored.

- **Lack of Comparative Studies Across ML Techniques:**

Most studies focus on a single ML approach (e.g., deep learning or clustering) without systematically comparing different models (e.g., KMeans vs. LightGBM vs. Neural Networks) under consistent backtesting conditions.

- **Insufficient Backtesting with Rebalancing:** Many existing works either neglect rebalancing or conduct simplified simulations that do not

reflect realistic trading conditions. Robust yearly rebalancing integrated with backtesting is not commonly applied across ML-based ESG portfolio strategies.

- **Underutilization of Sectoral and Structural Patterns:**

Sector-wise clustering or grouping is often overlooked, yet it can significantly enhance diversification and stability, especially when combined with ESG filtering.

- **Absence of AI-Driven Dynamic Portfolio Rebalancing Strategies:**

Static allocation models fail to adapt to evolving market and ESG conditions. A significant gap exists in applying ML-driven strategies for dynamic yearly rebalancing that can adjust to market volatility and shifts in ESG rankings.

## 3.2 Problem Statement

Long-term investing is essential for wealth creation, and with increasing global environmental concerns—such as pollution and climate change—sustainable investing has gained prominence. ESG (Environmental, Social, and Governance) scores now serve as key indicators for evaluating companies beyond financial metrics, enabling investors to align their portfolios with ethical and long-term goals.

Despite growing interest, many existing portfolio strategies either rely solely on traditional financial models or use machine learning (ML) without fully integrating ESG data. Moreover, they often neglect dynamic rebalancing and lack realistic long-term performance evaluation. Overdependence on volatile return predictions further limits their effectiveness.

This project proposes a comprehensive ML-based portfolio optimization framework that integrates ESG and financial features. Instead of relying on prediction alone, we backtest performance using real stock data (2021–2025) and benchmark it against the S&P 500 for realistic evaluation.

Key contributions include:

- **ESG Integration:** Combining ESG scores with financial indicators for data-driven stock selection and allocation.
- **Multi-Model Strategy:** Utilizing K-Means clustering with Mean-Variance Optimization (MVO), neural networks, and LightGBM Ranker to diversify methods.
- **Dynamic Rebalancing:** Annual rebalancing to reflect changing ESG scores, market trends, and model outputs.
- **Backtesting Focus:** Emphasizing historical performance over return

prediction for more reliable evaluation.

This approach aims to build adaptive, sustainable, and high-performing portfolios aligned with evolving market and environmental dynamics.

# Chapter 4

## Proposed Methodology/ Solution

### 4.1 Dataset Overview

To construct and evaluate our sustainable portfolio strategies, we utilized four primary datasets:

- **ESG Dataset (`esg_2021.csv`)**

1) Contains Environment, Social and Governance (ESG) scores for S&P 500 companies as of 2021.

2) *Key features:* `totalEsg`, `esgPerformance`, `highestControversy`, `socialScore`, `governanceScore`, `environmentalScore` etc. used for incorporating sustainability metrics into the stock selection process.

3) These were obtained from a curated GitHub repository compiling publicly available ESG metrics.

- **Financial Dataset (`filtered_security_data.csv`)**

**Purpose:** To inform value, quality, and liquidity-based stock selection and ranking in model-based strategies.

**Why used:** Financial ratios help in identifying undervalued yet fundamentally strong companies. They also support traditional factor investing strategies.

*Key features:* averageVolume, trailingPE, priceToBook, beta, profitMargins, etc

- **Daily stock prices data (snp500\_stocks\_closing\_price\_daily\_data.csv)**

1) Daily closing prices of individual S&P 500 stocks from Mar 2021 to Feb 2025.

2) Used for computing past returns, calculating portfolio returns and backtesting strategies.

3) This data were retrieved from yahoo finance.

- **S&P 500 Index data (snp500\_INDEX\_daily\_closing\_prices.csv)**

1) Daily closing prices of the S&P 500 index over the same period.

2) Serves as a benchmark to evaluate the relative performance (e.g. CAGR comparison) of the constructed portfolio.

3) This data were also retrieved from yahoo finance.

## 4.2 Data Preprocessing

To prepare the data for analysis, the following preprocessing steps were applied:

1. Merged ESG and financial datasets using company tickers.
2. Removed non-numeric and irrelevant numeric data that was not important for the analysis.
3. Standardized all financial indicators using Z-score normalization to create composite investment factors:

- **Value Factor:** Negative Z-scores 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.
4. Filtered out stocks with low trading liquidity (less than 1 million average daily volume).
  5. Retained only firms with above-median ESG scores to ensure a sustainable investment universe.
  6. Created ESG data for 2022, 2023, 2024 and 2025 using random variations from the previous year's ESG scores. This method simulates potential shifts in ESG rankings and ensures an extended time horizon for the portfolio optimization process. The new ESG datasets were saved and utilized for the model's prediction and optimization steps.

### 4.3 Baseline Strategy: KMeans Clustering

- **Feature Preparation:** ESG and financial metrics (e.g., `totalEsg`, `trailingPE`, `priceToBook`, `profitMargins`) from 2021 were normalized using z-scores to ensure consistency across features.
- **Two-Layered K-Means Clustering:** To ensure diversification and ESG alignment, we implemented a hierarchical clustering strategy using two successive KMeans clustering layers:

#### 1. First-Layer Clustering: Financial-Based Grouping

- Stocks were clustered based on standardized financial factors covering value, quality, momentum, and growth dimensions.



- Sectoral constraints were incorporated to preserve industry-level diversification.
- The top-performing financial cluster was selected based on the average composite score, which aggregated the financial metrics.

## 2. Second-Layer Clustering: ESG-Based Refinement

- Stocks within the selected financial cluster were reclustered based solely on ESG-related metrics.
  - The ESG-compliant cluster with the highest average ESG score was selected for final stock selection.
- **Stock Selection:** From the final ESG-refined cluster, the top 20 stocks were selected based on their composite score (integrating ESG, value, and quality metrics).
  - **Portfolio Construction:** Portfolio weights were then optimized using Mean-Variance Optimization (MVO), aiming to maximize expected return minus half the portfolio variance. The resulting weights were held constant over the full backtesting period from 2021 to 2025.
  - **Evaluation:** Daily returns were calculated using the selected stocks price data, and the portfolio's cumulative return was computed over time. This performance was then benchmarked against the cumulative return of the S&P 500 index over the same period to assess relative performance. The Compound Annual Growth Rate (CAGR) was used as a key metric for comparison.
  - **Limitations:**
    - 1) The static nature of the model does not adapt to evolving market or ESG dynamics.

2) Time-series dependencies and momentum effects are ignored, potentially reducing forecasting strength.

## 4.4 LightGBM-Based Portfolio Construction

The LightGBM model uses the current year’s ESG and financial data to predict the future returns of each stock. The methodology for constructing the portfolio is as follows:

- **Data Merging** For each rebalancing year (2022, 2023, 2024), the dataset is merged with the previous year’s data to calculate the ESG Change (`esg_change`).
- **Target Calculation** Future returns are computed using the price data. For each stock, the percentage change in its price from the beginning of the year to the end is calculated.
- **Feature Set** The features include ESG data (`totalEsg` and `esg_change`), financial metrics (e.g., `priceToBook`, `trailingPE`), and other relevant information, while the target is the calculated forward return.
- **Model Training** The LightGBM model is trained on the current year’s data to predict the future returns. Hyperparameters such as learning rate, number of leaves, and the objective function are set for regression tasks.
- **Prediction and Ranking** Once the model is trained, it predicts the returns for all available stocks. The stocks are then sorted based on their predicted return (or, if `esg_change` is available, an ESG-weighted predicted return). The top 20 stocks are selected for the portfolio.

## 4.5 Neural Network-Based Portfolio Construction

The Neural Network approach focuses on using ESG scores and financial data to predict future returns through an MLP (Multi-Layer Perceptron) model:

- **Data Merging and Feature Scaling** ESG and financial data for each year is merged and cleaned. Non-numeric columns are dropped, and the remaining numeric data is standardized using `StandardScaler` to ensure that the features are on the same scale for the neural network model.
- **Forward Returns Calculation** The 1-year forward returns for each stock are calculated by comparing the price at the start and end of the year.
- **Training the Model** The neural network is trained on the past data for each year using PyTorch. The model consists of an input layer, two hidden layers with ReLU activations, and an output layer that predicts the stock's return. The model is trained using mean squared error (MSE) loss and the Adam optimizer for 100 epochs.
- **Prediction and Ranking** The trained model is used to predict the returns of stocks for the subsequent year (e.g., predicting 2024 returns based on 2023 data). The stocks are then sorted based on their predicted returns, and the top 20 stocks are selected.

## 4.6 Portfolio Construction and Evaluation

For both the LightGBM and Neural Network methods, the top 20 stocks are selected each year. The portfolio weighting can be done in two ways:

- **Equal Weights** In both models, the portfolio is equally weighted, i.e., each of the selected 20 stocks has the same weight in the portfolio.
- **ESG-Weighted Portfolio** In the LightGBM model, an optional adjustment is made to the predicted returns by applying a small boost based on the ESG Change of each stock.

Once the stocks are selected, the cumulative returns for the portfolio are calculated by tracking the daily price movements of the selected stocks.

## 4.7 Backtesting and Performance Evaluation

The portfolio is backtested by simulating its performance over time, from March 2021 to February 2025:

- **Normalized Returns** The portfolio's returns are normalized to a starting value of 1 to show relative performance over time.
- **Portfolio Returns Calculation** For each stock in the portfolio, its price data is fetched, and its normalized return is calculated. The portfolio returns are calculated as the sum of the normalized returns of the top stocks.
- **CAGR Calculation** The Compound Annual Growth Rate (CAGR) is calculated to evaluate the performance of the portfolio over the entire period. The CAGR is used as a key metric to assess the effectiveness of the portfolio strategy.

# Chapter 5

## Experimental Setup

### 5.1 Datasets

1)Esg dataset:-

symbol	governancescore	environmentalScore	totalesg	esgPerformance	percentile
ABT	10.65	2.98	29.85	AVG_PERF	56.59
ANSS	4.84	1.13	14.94	UNDER_PERF	14.94
GOOGL	11.94	0.48	22.29	AVG_PERF	28.96
PPL	5.47	18.72	35.17	LEAD_PERF	74.34
NVDA	6.88	2.33	13.37	LAG_PERF	4.83
AAPL	8.94	0.1	16.73	UNDER_PERF	11.77
AMZN	9.78	5.13	27.42	AVG_PERF	47.67

Table 5.1: ESG data

## 2)Financial Data:-

symbol	averageVolume	priceToBook	trailingPE	profitMargins	Beta
ABT	4984786	6.2607374	46.296444	0.12988	0.723185
ANSS	763757	6.5509667	62.392357	0.25807	1.215059
GOOGL	1845114	6.1780233	34.739563	0.22062	0.999458
PPL	6493700	1.5788867	14.376963	0.19311	0.783315
NVDA	11314814	18.304033	72.27971	0.25979	1.412242
AAPL	134912557	30.482723	32.541363	0.21735	1.251354
AMZN	4047200	16.466013	73.09682	0.05525	1.132719

Table 5.2: Financial Data

## 3)S&P500 stocks closing price daily data:-

symbol	1 Mar 2021	2 Mar 2021	3 Mar 2021	26 Feb 2025	27 Feb 2025
ABT	122.21	122.53	119.18	135.96	135.97
ANSS	346.11	336.33	319.86	332.32	330.37
GOOGL	103.48	103.22	100.57	170.28	168.5
PPL	26.89	27.14	27.48	34.87	34.45
NVDA	13.84	13.41	12.8	131.28	120.15
AAPL	127.79	125.12	122.06	240.36	237.3
AMZN	157.31	154.73	150.25	214.35	208.74

Table 5.3: Daily Prices data of S&P500 stocks

#### 4) S&P500 INDEX Daily Closing Prices

Date	Closing Price
1 Mar 2021	3901.82
2 Mar 2021	3870.29
3 Mar 2021	3819.72
26 Feb 2025	5956.06
27 Feb 2025	5861.57
28 Feb 2025	5954.5

Table 5.4: S&P500 INDEX Daily Closing Prices

## 5.2 Parameter Settings

To ensure reproducibility and transparency, the key hyperparameters and simulation configurations for each model are listed below.

- **Portfolio size:** Top 20 stocks
- **Rebalancing frequency:** Yearly

### LightGBM Ranker

- Objective: `regression`
- Boosting Type: `gbdt` (Gradient-Boosted Decision Trees)
- Learning rate: 0.05

### Neural Network

- Architecture:
  - Dense(128) + ReLU + Dropout(0.2)
  - Dense(64) + ReLU
  - Dense(1) output

- Loss function: MSE (Mean Squared Error)
- Optimizer: Adam (Adaptive Moment Estimation)
- Epochs: 100
- Batch size: 32

## 5.3 Hardware and Software Requirements

- **Hardware:**

- CPU: Intel i5 / AMD Ryzen 7 or equivalent
- RAM: Minimum 16 GB

- **Software:**

- Operating System: Ubuntu 22.04 / Windows 11
- Programming Language: Python 3.x+
- Libraries: pandas, numpy, tensorflow, lightgbm, matplotlib, scikit-learn, scipy



## 5.4 Simulation Flow

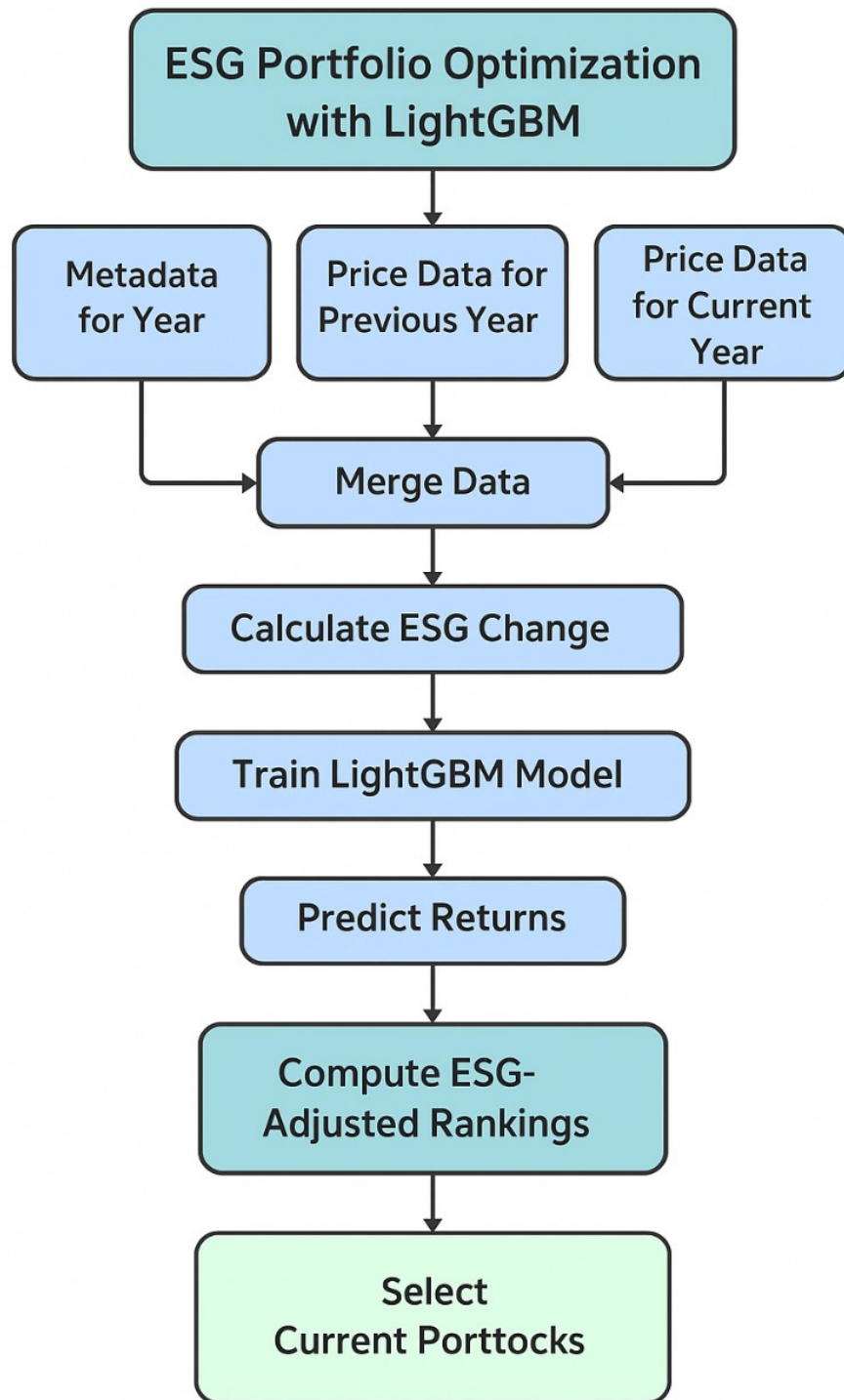


Figure 5.1: Flowchart of LighGBM Process

## 5.5 Evaluation Metrics

To assess the performance of the portfolio strategies, the following evaluation metrics were used:

### Compound Annual Growth Rate (CAGR)

CAGR measures the mean annual growth rate of an investment over a specified period of time longer than one year. It is calculated as:

$$CAGR = \left( \frac{V_f}{V_i} \right)^{\frac{1}{T}} - 1 \quad (5.1)$$

where:

- $V_f$  is the final portfolio value
- $V_i$  is the initial portfolio value
- $T$  is the duration of the investment in years

### Sharpe Ratio

Sharpe Ratio evaluates the return of an investment compared to its risk. It is calculated as:

$$SharpeRatio = \frac{E[R_p - R_f]}{\sigma_p} \quad (5.2)$$

where:

- $R_p$  is the portfolio return
- $R_f$  is the risk-free return (assumed to be 0% in this analysis)
- $\sigma_p$  is the standard deviation (volatility) of the portfolio return

## **Benchmark Comparison**

The performance of the portfolio was also compared to the S&P 500 index by calculating its CAGR over the same evaluation period (2021–2025).

# Chapter 6

## Results and Discussion

### 6.1 Portfolio Performance Analysis

- The portfolio optimization was conducted using three distinct models: **LightGBM Ranker**, **KMeans Clustering**, and **Neural Network**. All models incorporated ESG factors and financial metrics, and were backtested against the S&P 500 index from 2021 to 2025. The Mean-Variance Optimization (MVO) was used for portfolio optimization, and the performance was evaluated based on key metrics such as CAGR and Sharpe ratio.
- The results demonstrate that each model had a unique impact on portfolio performance. The LightGBM model achieved the highest CAGR, followed by the KMeans clustering model. The Neural Network model has low returns but outperforms benchmark.
- The LightGBM model achieved a CAGR of 40.10%, significantly outperforming the S&P 500 benchmark (CAGR = 11.42%). The KMeans model achieved a CAGR of 27.83%, while the Neural Network model achieved a CAGR of 13.94%.

## 6.2 Performances of Models:-

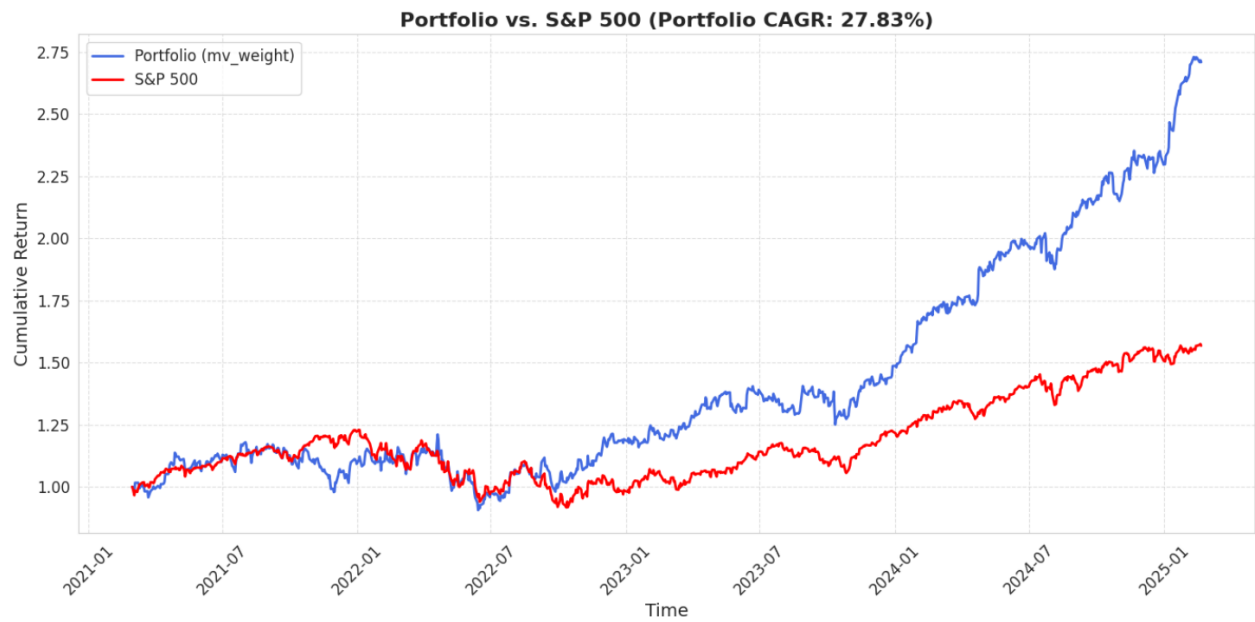


Figure 6.1: Comparison of Portfolio constructed using Kmeans Clustering with S&P 500

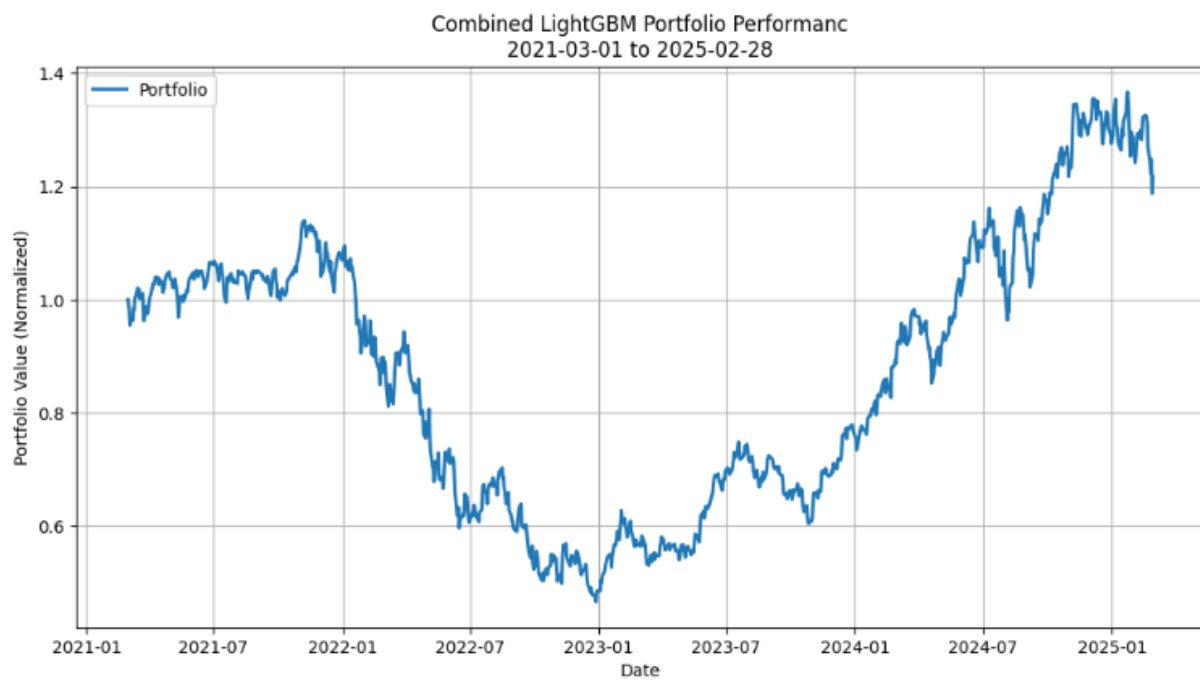


Figure 6.2: Comparison of Portfolio constructed using LightGBM Ranker

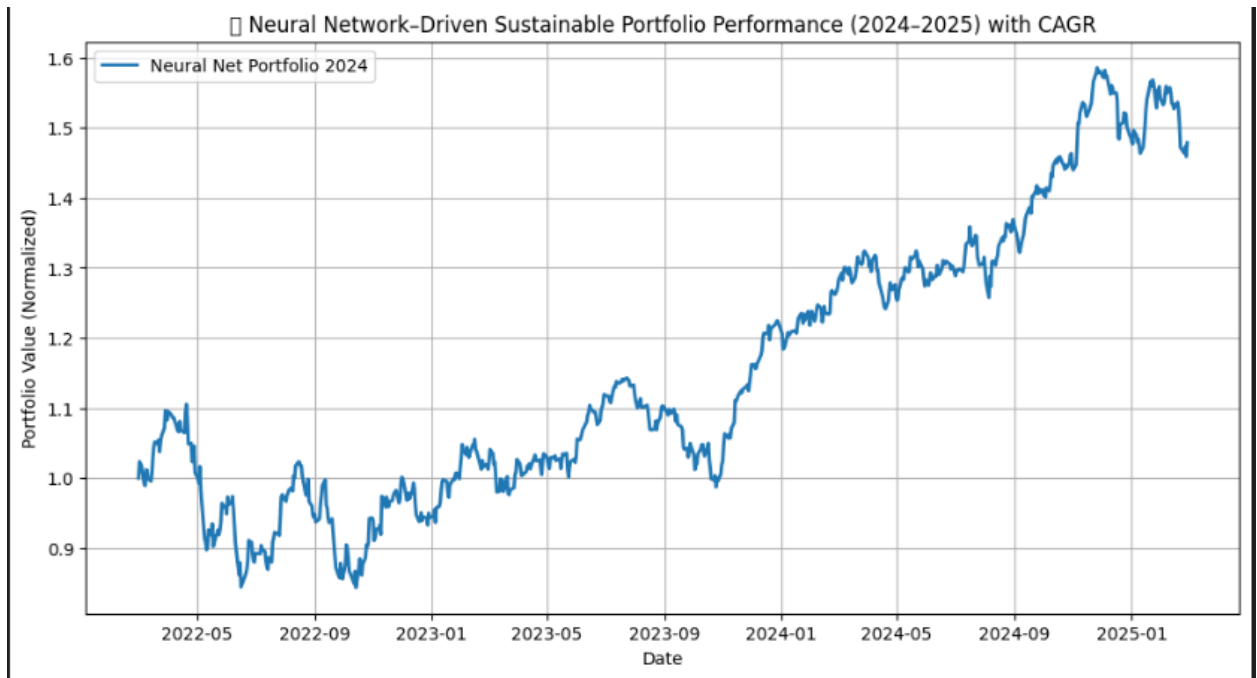


Figure 6.3: Comparison of Portfolio constructed using Neural Network

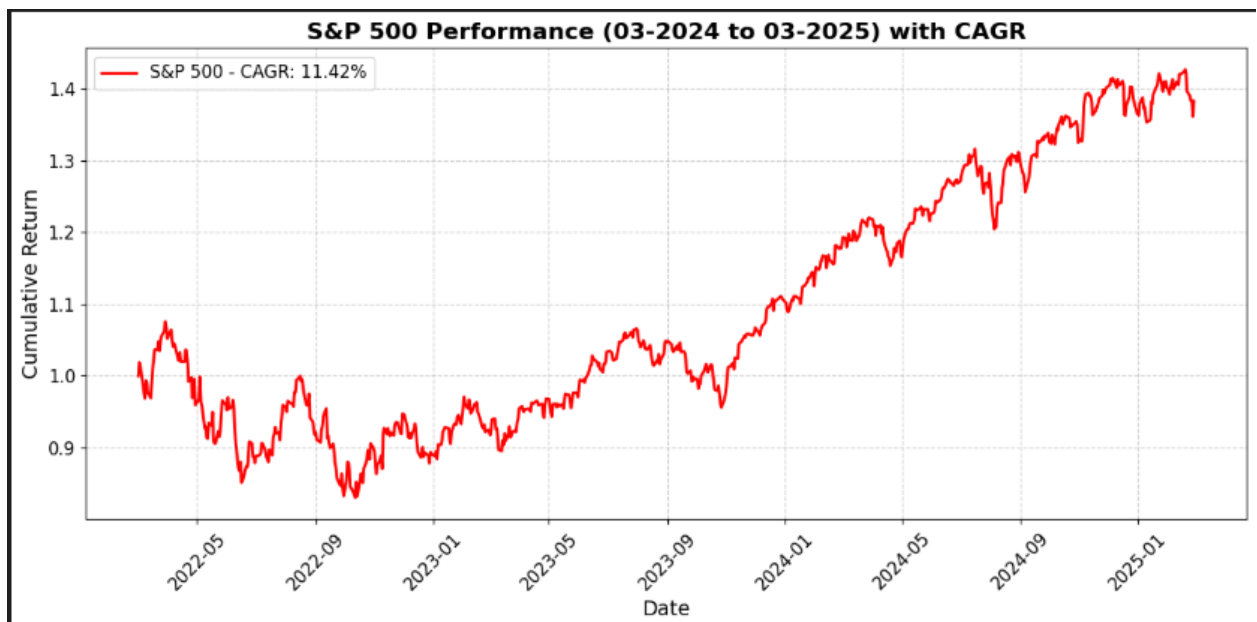


Figure 6.4: S&P 500 Index Graph

## 6.3 Key Observations

- **LightGBM Ranker Model:** Delivered the highest Compound Annual Growth Rate (CAGR) of 40.10%, outperforming both the S&P 500 benchmark and the other two models. This highlights the strong predic-

tive power of gradient boosting models in integrating ESG and financial data for effective portfolio construction.

- **KMeans Clustering Model:** Achieved a solid CAGR of 27.83%, offering good diversification and ESG alignment. While not as high-performing as LightGBM, its rule-based structure adds robustness and interpretability to portfolio selection.
- **Neural Network Model:** Acheived a lower CAGR of 13.94% compared to other models, indicating that despite its flexibility, the model may require more sophisticated architecture or more data to effectively capture financial patterns in this context.
- **S&P 500 Benchmark:** Delivered a CAGR of 11.42% during the same period, demonstrating the potential benefits of applying ESG-integrated, data-driven strategies over traditional passive investing.

Model	CAGR (%)
LightGBM Ranker	40.10
KMeans Clustering	27.83
Neural Network	13.94
S&P 500 Benchmark	11.42

Table 6.1: Portfolio Performance Comparison (2021–2025)

# Chapter 7

## Conclusion

### 7.1 Summary of Contributions

This research presents a comprehensive approach to sustainable portfolio optimization by integrating Environmental, Social, and Governance (ESG) metrics with traditional financial indicators. Three modeling strategies—KMeans Clustering, LightGBM Ranker, and Neural Networks—were implemented and compared over the period 2021–2025. A unified backtesting framework with yearly rebalancing ensured consistent evaluation across all models.

The primary contributions of this study are:

- A unified methodology for incorporating ESG and financial features into machine learning-based portfolio construction.
- Development and benchmarking of unsupervised and supervised learning models (KMeans, LightGBM Ranker, Neural Network) for sustainable investing.
- Implementation of a yearly rebalancing backtesting system based on historical daily stock price data.
- Performance comparison with the S&P 500 index to assess the viability of intelligent, ESG-aligned strategies.



## 7.2 Key Findings

- The **LightGBM Ranker model** achieved the highest CAGR of **40.10%**, demonstrating the effectiveness of gradient boosting algorithms in leveraging ESG and financial data for stock ranking and portfolio construction.
- The **KMeans Clustering model** delivered a moderate CAGR of **27.83%**, showing that unsupervised clustering can be useful for diversification, though less effective in return maximization.
- The **Neural Network model** recorded the lowest CAGR of **13.94%**, indicating the need for further tuning or more sophisticated architectures to fully capitalize on its learning capacity.
- The **S&P 500 benchmark** achieved a CAGR of **11.42%**, significantly underperforming compared to all the model-driven portfolios, highlighting the advantage of data-driven ESG investment strategies.

## 7.3 Limitations

Despite promising results, several limitations should be noted:

- ESG scores for 2022, 2023 and 2024 were synthetically manipulated based on 2021 data, which may not accurately reflect the true evolution of corporate sustainability over time.
- The study does not incorporate macroeconomic indicators or behavioral factors, which can significantly influence financial markets and stock returns.
- The black-box nature of neural networks reduces model interpretability, which is a crucial consideration in responsible and transparent ESG

investing.

## 7.4 Future Work

Future research can improve upon this study in the following ways:

- Utilizing real, time-varying ESG datasets across multiple years to more accurately capture the dynamic sustainability profiles of companies, instead of relying on manipulated or static ESG data.
- Incorporating dynamic financial metrics instead of using only static 2021 financial data, to better reflect evolving company fundamentals over time.
- Enhancing performance evaluation through additional risk-adjusted metrics such as Sharpe Ratio, Sortino Ratio, and Maximum Drawdown.
- Investigating advanced and adaptive modeling techniques such as Reinforcement Learning or hybrid approaches (e.g., KMeans combined with LightGBM) for robust and context-aware portfolio construction.
- Incorporating real-world constraints like transaction costs, liquidity constraints, capital gains taxation, and market impact for more practical and implementable investment strategies.
- Applying Explainable AI (XAI) frameworks to improve the transparency and interpretability of model predictions, which is especially important in ESG-focused and responsible investment strategies.

# Bibliography

- [1] H. N. Bhandari, N. R. Pokhrel, R. Rimal, et al. Implementation of deep learning models in predicting esg index volatility. *Financial Innovation*, 10:75, 2024.
- [2] L. Chan and T. Seah. Artificial neural networks in portfolio optimization. *Quantitative Finance Review*, 2024.
- [3] Carmine de Franco, Christophe Geissler, Vincent Margot, and Bruno Monnier. Esg investments: Filtering versus machine learning approaches. *The Seventh Public Investors Conference*, 2018.
- [4] X. Feng, H. J. Mettenheim, G. Sermpinis, and C. Stasinakis. Sustainable portfolio construction via machine learning: Esg, sdg and sentiment. *European Financial Management*, 2024.
- [5] Eduardo C. Garrido-Merchán, Sol Mora-Figueroa-Cruz-Guzmán, and María Coronado-Vaca. Deep reinforcement learning for esg financial portfolio management. *arXiv preprint*, 2023.
- [6] T. Lim. Environmental, social, and governance (esg) and artificial intelligence in finance: State-of-the-art and research takeaways. *Artificial Intelligence Review*, 57:76, 2024.
- [7] Sean M. Lynch. Artificial intelligence in stock analysis and portfolio building, 2024.

- [8] V. Maree and S. Omlin. Reinforcement learning for esg portfolio optimization. *Journal of Machine Learning in Finance*, 2024.
- [9] Alejandro Momparler, Pablo Carmona, and Francisco Climent. Catalyzing sustainable investment: Revealing esg power in predicting fund performance with machine learning. *Computational Economics*, 65:1617–1642, 2025.
- [10] Taeisha Nundlall and Terence L. Van Zyl. Machine learning for socially responsible portfolio optimisation. *arXiv preprint*, 2023.
- [11] 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*, 8(2):24–36, 2024.
- [12] Michael Schopf. Advancing portfolio construction and optimization: Ai’s role in boosting returns, lowering risks, and streamlining efficiency. *SSRN*, 2024.
- [13] Kamurthi Ravi Teja and Chuan-Ming 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*, 2024:1–XX, 2024.
- [14] Jun Xu. Ai in esg for financial institutions: An industrial survey. *SSRN*, 2024.

INTERNATIONAL CONFERENCE ON  
BIG DATA AND ARTIFICIAL INTELLIGENCE AND IoT(ICBDAIT)

## Certificate OF PRESENTATION

This is to certify that

**Dr. P. R. Deshmukh**

has presented a research paper entitled “Sustainable Investing with  
AI: K-Means Clustering, MVO and Back Testing Approach for  
Portfolio Construction” at the International Conference on Big  
Data and Artificial Intelligence and IoT (ICBDAIT) held in  
Pune, India on 28<sup>th</sup> April, 2025.



  
CONFERENCE COORDINATOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)

  
  
MANAGING DIRECTOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)

PAPER ID

AR-DAIT-PUNE-280425 - 12752

WWW.ARSSS.ORG

INTERNATIONAL CONFERENCE ON  
BIG DATA AND ARTIFICIAL INTELLIGENCE AND IoT(ICBDAIT)

## Certificate OF PRESENTATION


This is to certify that

**Nachiket Deshmukh**

has presented a research paper entitled “Sustainable Investing with  
AI: K-Means Clustering, MVO and Back Testing Approach for  
Portfolio Construction” at the International Conference on Big  
Data and Artificial Intelligence and IoT (ICBDAIT) held in  
Pune, India on 28<sup>th</sup> April, 2025.



  
CONFERENCE COORDINATOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)

  
  
MANAGING DIRECTOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)

PAPER ID

AR-DAIT-PUNE-280425 - 12752

WWW.ARSSS.ORG

INTERNATIONAL CONFERENCE ON  
BIG DATA AND ARTIFICIAL INTELLIGENCE AND IoT(ICBDAIT)

## Certificate OF PRESENTATION

This is to certify that

**Ved Garudkar**

has presented a research paper entitled “Sustainable Investing with  
AI: K-Means Clustering, MVO and Back Testing Approach for  
Portfolio Construction” at the International Conference on Big  
Data and Artificial Intelligence and IoT (ICBDAIT) held in  
Pune, India on 28<sup>th</sup> April, 2025.



CONFERENCE COORDINATOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)



MANAGING DIRECTOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)

PAPER ID

AR-DAIT-PUNE-280425 - 12752

WWW.ARSSS.ORG

INTERNATIONAL CONFERENCE ON  
BIG DATA AND ARTIFICIAL INTELLIGENCE AND IoT(ICBDAIT)

## Certificate OF PRESENTATION

This is to certify that

**Nikhil Kokale**

has presented a research paper entitled “Sustainable Investing with  
AI: K-Means Clustering, MVO and Back Testing Approach for  
Portfolio Construction” at the International Conference on Big  
Data and Artificial Intelligence and IoT (ICBDAIT) held in  
Pune, India on 28<sup>th</sup> April, 2025.



CONFERENCE COORDINATOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)



MANAGING DIRECTOR  
ADVANCED RESEARCH SOCIETY FOR  
SCIENCE AND SOCIOLOGY  
(ARSSS)

PAPER ID

AR-DAIT-PUNE-280425 - 12752

WWW.ARSSS.ORG