International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Analysis of Clustering Methods for Portfolio Optimization

Achuta Mukund Harsha*[1], V.V.S. Kesava Rao[2]

Submitted: 23/10/2023 **Revised:** 14/12/2023 **Accepted:** 20/12/2023

Abstract: This paper presents a comprehensive exploratory analysis of financial performance of Indian companies, employing diverse clustering techniques and visualization methods. Core financial metrics, including Net Profit Margin (%), Earnings Per Share (EPS) Growth Rate (%), and Book Value (BV) Growth Rate (%), are examined over a decade. The study investigates optimal cluster numbers through Silhouette analysis and the Elbow Curve method, harnessing the power of unsupervised Machine Learning (ML). Employing K-Means and hierarchical clustering with various linkage strategies (single, complete, average, centroid, and ward), the research unveils insightful clustering structures through dendrogram visualizations. This analysis involves 50 undervalued companies, scrutinizing their financial dynamics and clustering profiles. The ensuing impact of indicators on trading activities and returns over both short and long time horizons (365 days and 10 years, respectively) is meticulously dissected. In particular, this study contrasts trade counts, total returns, and average return per trade vis-à-vis a benchmark of Nifty50 companies. The research findings substantiate that a majority of the clusters yield higher returns compared to Nifty50 counterparts for identical technical indicator combinations. The application of advanced ML methodologies in this research provides actionable insights from complex financial data, catering to the needs of both researchers and practitioners. Ultimately, this work enriches financial analysis and offers valuable inputs for refining trading strategies.

Keywords: *Clustering Techniques; Machine Learning; Portfolio Optimization; Fundamental Analysis; Indian Financial Markets.*
JEL Classification codes: *C38, C53, G11; G17.*

1. Introduction

In the dynamic landscape of financial analysis and investment strategies, data-driven techniques have emerged as powerful tools to navigate the complexities of modern markets. Machine Learning (ML), a subfield of artificial intelligence, has gained prominence for its ability to uncover hidden patterns and insights within vast datasets. The fundamental challenge lies in efficiently extracting actionable insights from complex financial data, a task that traditional methods often struggle to accomplish. Here, the application of ML techniques provides a transformative approach. By autonomously identifying inherent structures and relationships within datasets, these techniques enable the dissection of intricate financial behaviours that might otherwise remain obscured. An extensive analysis of Indian companies' financial performance, focused on core financial metrics such as Net Profit Margin (%), Earnings Per Share (EPS) Growth Rate (%), and Book Value (BV) Growth Rate (%), studied over a decade. This analysis showcases the potential of ML, as exemplified by Silhouette analysis and the Elbow Curve method, to uncover optimal cluster numbers and patterns within the data.

The utilization of K-Means and hierarchical clustering, incorporating various linkage strategies, unveils insightful clustering structures through dendrogram visualizations. The implications of this analysis are far-reaching, extending beyond clustering profiles to explore the effects of technical indicators on trading activities and returns. By scrutinizing a span of 365 days for short-term analysis and a decade for long-term evaluation, the study contrasts trade counts, total returns, and average return per trade with Nifty50 benchmark companies.

This research underscores the potential of advanced ML methodologies and visualization tools as invaluable assets within the domain of financial analysis and decision support systems. The findings presented in this paper not only enrich understanding of financial dynamics but also offer actionable insights that can drive informed investment strategies and decision-making processes in complex market environments.

2. Literature Review:

This collection of works spans various facets of portfolio optimization and investment strategies. Reference [1] highlights the superiority of ML models over conventional methods in predicting stock returns during earnings announcements, providing valuable insights for value-oriented investors. Study [2] deepens this understanding by exploring ML techniques within Chinese value investing, revealing their potential to create more effective strategies based on value-related indicators. In reference [3], ML's integration into professional investors' decision-making processes leads to anomaly-driven strategies in stock market investments.

The synergy between fundamental principles and ML techniques shines in reference [4], enhancing value-centered investment strategies. The authors of reference [5] assert that AI-driven value strategies yield superior risk-adjusted returns in the US stock market. Study [6] introduces an innovative ML framework for value-oriented investing, incorporating deep learning, transfer

[1]*Department of Mechanical Engineering, Andhra University, Visakhapatnam, India,*
ORCID ID: 0000-0001-5538-7765
[2] *Faculty of Industrial Engineering, Department of Mechanical Engineering, Andhra University, Visakhapatnam, India,*
ORCID ID: 0000-0002-0905-9688
** Corresponding Author Email: mukundharsha.rs@andhrauniversity.edu.in*

learning, and reinforcement learning, ultimately boosting accuracy in identifying valuable stocks.

Dynamic portfolio optimization strategies take center stage in multiple studies. Reference [7] presents an adaptable approach based on evolving Minimum Spanning Tree structures in Chinese stock markets, adjusting to market conditions for optimal outcomes. On the other hand, [8] employs non-negative matrix factorization (NMF) to cluster stocks based on underlying factors.

Cluster analysis proves crucial for portfolio diversification strategies across diverse works. Reference [9] introduces novel financial ratio-based similarity measures for cluster analysis, excelling in challenging market periods. Similarly, [10] leverages cluster analysis and Sharpe ratios to craft effective investment portfolios, outperforming random selection with better volatility performance.

The interplay between innovation and investment strategies unfolds in [11], investigating firm cluster collaboration and big data's impact on open innovation expansion. Positive associations are observed, though social capital's moderating influence is noted. Notable portfolio optimization methodologies include [12]'s use of technical indicators and investor attitudes, optimizing portfolios through fuzzy numbers and a genetic algorithm. [13] employs data mining and clustering for risk-minimizing diversification, effectively preserving returns.

Innovation-driven clustering approaches emerge in [14], blending hierarchical clustering with stock price momentum for enhanced returns and portfolio stability. [15] adds a unique perspective by utilizing valuation ratios to identify clusters among Indian small-cap companies, contributing to unconventional diversification.

The interconnectedness of portfolios and risk analysis takes the forefront in [16], introducing a portfolio volatility spillover index. This index underscores increased interconnectedness amid portfolios, shaped by foreign exchange markets, necessitating continuous monitoring for effective management. [17] focuses on adaptable portfolio strategies by clustering market states using historical data, leading to higher Sharpe Ratios and resilience. [18]'s "minCluster portfolio" approach combines downside risk, hierarchical clustering, and robustness, excelling in volatile markets.

Innovative portfolio construction methods are showcased in [19], utilizing shape-based time-series clustering for enhanced diversification. This method outperforms traditional approaches. Additionally, [20] emphasizes balanced portfolio construction in smaller markets like the Abu Dhabi Securities Exchange (ADX), while [21] personalizes portfolios based on investor risk preferences, offering fresh perspectives.

Research on portfolio volatilities and spillover effects is deeply examined in [16], introducing a portfolio volatility spillover index that outshines traditional indicators. The interconnectedness of portfolios, evolved since the Global Financial Crisis, underscores continuous monitoring. Innovative portfolio diversification remains central to [22], introducing a novel method based on shape-based time-series clustering that bolsters risk-adjusted returns.

The integration of data mining and granular computing for portfolio optimization is exemplified in [23], successfully tested on Hong Kong Stock Exchange-listed stocks. Lastly, [24] introduces a distinctive approach for novice investors, utilizing multicriteria decision-making techniques in a fuzzy environment. This method categorizes assets and constructs portfolios using a fractional lion clustering algorithm, outperforming major Index funds in returns while maintaining comparable risk, thereby redefining portfolio construction strategies.

Literature Gaps Identified:

Based on the literature review, there appear to be several potential gaps or areas where this research can make a unique contribution:

1. Integration of ML with Fundamental Analysis: While ML techniques have shown promise in Western markets, their integration with traditional fundamental analysis remains underexplored in the Indian context. The gap lies in comprehending how these two approaches can synergize effectively to enhance investment decisions.

2. Optimal Clustering Techniques: The choice of clustering methods for portfolio diversification in the Indian market is not well-established. A gap exists in identifying which linkage methods and distance metrics work best for clustering Indian stocks to maximize portfolio diversification.

3. Performance Evaluation in Indian Context: There's a need to assess the performance of ML-driven portfolio strategies in the Indian market comprehensively. This gap pertains to understanding how these strategies fare in terms of risk-adjusted returns, especially when applied to Indian stocks.

4. Practical Application for Indian Investors: A gap exists in providing practical guidance and insights for Indian investors on how to leverage ML-based portfolio optimization effectively, taking into consideration the specific characteristics and constraints of the Indian stock market.

5. Holistic Risk Management: Many studies touch upon risk analysis, but a gap remains in developing holistic risk management strategies tailored to the Indian market, considering factors such as interconnectedness and volatility.

Objectives of the Study:

1. To develop a customized ML-driven portfolio optimization methodology tailored to the Indian stock market's unique characteristics and challenges.

2. To assess the integration of traditional fundamental analysis with ML techniques, showcasing how this fusion can enhance stock valuations and investment decisions in the context of the Indian stock market.

3. To conduct a comprehensive analysis of clustering techniques, comparing various linkage methods and distance metrics, with a specific focus on optimizing portfolio diversification for Indian market investments.

4. To evaluate the performance of ML-driven portfolio strategies, particularly in terms of risk-adjusted returns, and determine their practical effectiveness for Indian investors.

3. Methodology

The methodology section of this research paper employs a comprehensive approach to unravel the intricate dynamics of the financial performance of Indian companies. To achieve a nuanced understanding, a diverse set of clustering techniques and visualization methods is applied, facilitating an in-depth

exploratory analysis.

3.1 Data Collection and Processing:

Building upon the prior research [25], the authors employ a data collection process to identify the top 50 undervalued companies. This process involves a Financial Decision Support System (DSS) that combines traditional fundamental analysis with ML. By utilizing Random Forest model, the DSS scores stock valuations, effectively refining options from a larger pool identified through stock screeners. Initially, 189 stocks were screened using online stock screeners, with 140 of them having a decade of historical data used as inputs for the ML model. Applied to Out of Time Data for the selected 140 stocks, the Random Forest model identified 57 stocks scoring above 80% out of which 15 stocks scoring above 90% as highly undervalued. Ultimately, the DSS effectively selects the most undervalued stocks based on their robust ML scores, offering a powerful tool for investment decision-making.

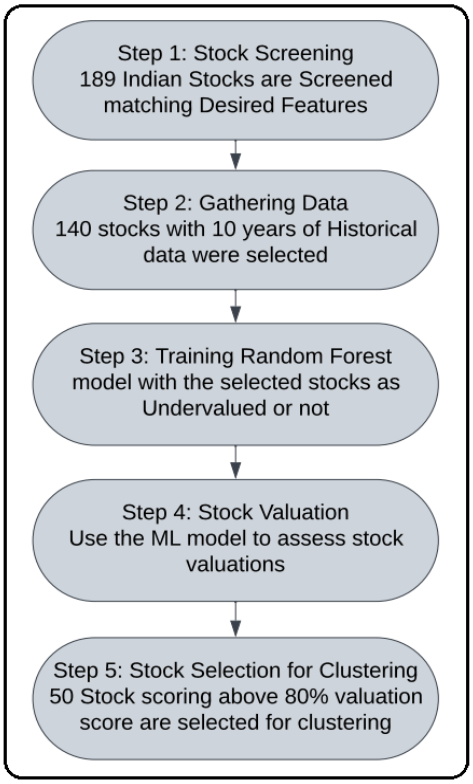

Fig. 1. Workflow of Financial-DSS for Stock Valuation

3.2 Unsupervised ML Technique:

Utilizing the ML scores derived from the Data Scoring System (DSS), this study delves into the utilization of unsupervised ML methodologies for the purpose of clustering a carefully chosen set of 50 undervalued companies. The study primarily focuses on fundamental financial indicators spanning a decade, notably encompassing metrics such as Net Profit Margin (%), Earnings Per Share (EPS), and Book Value (BV).

In order to standardize the data and facilitate a comprehensive analysis of stock performance, the variables used for clustering encompass Net Profit Margin (%), Growth Rate of Earnings Per Share (EPS) (%), and Growth Rate of Book Value (BV) (%). The resulting clustering aims to offer insights into the interrelationships and commonalities within this group of companies. For a clearer understanding of the dataset employed in the modelling process, refer to Table 1, which provides a concise overview of the company-specific data considered over the span of ten years.

Table 1. Sample historical Data of single Company

Financial Year	Basic EPS (Rs.)	Net Profit Margin (%)	Book Value (Rs.)	EPS Growth Rate (%)	BV Growth Rate (%)
Mar 14	2.98	3.44	16.02	20	20
Mar 15	2.75	2.79	18.79	-8	17
Mar 16	2.72	2.43	21.42	-1	14
Mar 17	4.4	3.99	26.7	62	25
Mar 18	4.8	4.04	38.61	9	45
Mar 19	3.96	3.53	44.85	-18	16
Mar 20	5.77	4.89	50	46	11
Mar 21	8.57	6.31	59.1	49	18
Mar 22	9.13	5.78	67.19	7	14
Mar 23	12.5	5.69	80.55	37	20

3.3 Assessment of Clustering Tendency and Identifying Optimal Number Clusters:

The assessment of clustering tendency is conducted utilizing the Hopkins statistic. This statistical method assesses clustering tendencies by contrasting the distribution of distances between randomly sampled dataset points (U-Distance) with the distribution of distances between data points and their nearest neighbors (W-Distance). A discernible difference between the U-Distance and W-Distance distributions indicates a clustering tendency. The Hopkins statistic yields values within the range of 0.95 to 0.99, indicative of a pronounced clustering tendency.

$$Hopkinsstatistic(H) = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} w_i} \quad (1)$$

Where:

'n' is the number of points in the dataset.

'ui' is the distance to the nearest neighbour in the artificial dataset for point.

'wi' is the distance to the nearest neighbour in the original dataset for point.

To determine the optimal number of clusters, a two-pronged approach involving Silhouette analysis and the Elbow Curve method is undertaken. The Elbow curve showcases the interplay between the Sum of Squared Distances and the number of clusters, aiding in identifying an inflection point. Meanwhile, the Silhouette analysis bestows Silhouette scores upon varying cluster numbers, gauging the quality of cluster assignments. Operating on a scale of -1 to 1, where 1 signifies dense and well-separated clusters, the Silhouette score plays a pivotal role in the decision-making process. Upon careful analysis, the optimal cluster count is ascertained to be 3. Figure 2, illustrates the dynamic interplay of the Elbow Curve and Silhouette scores across diverse cluster counts.

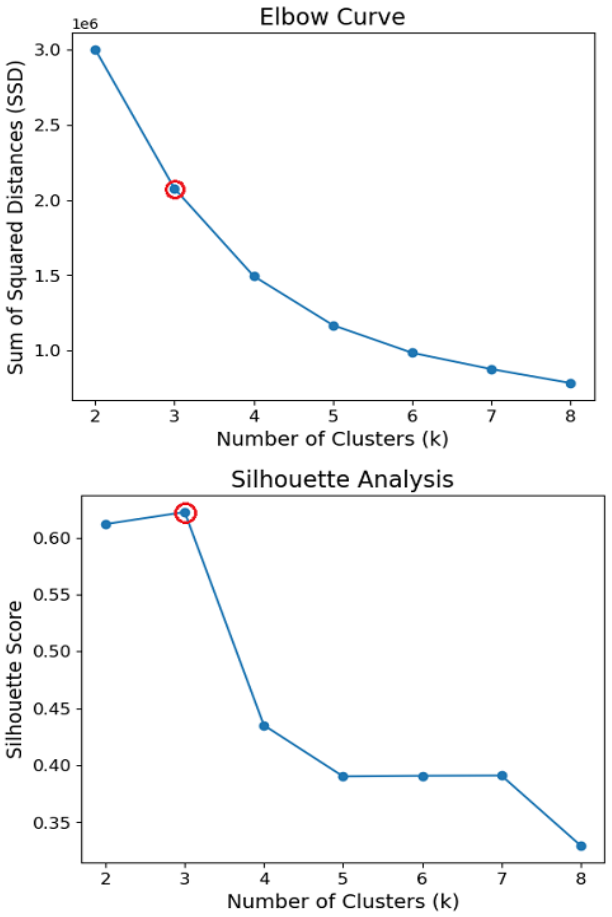

Fig. 2. Optimal Number of Cluster using Elbow curve and Silhouette Analysis

3.4 Clustering and Visualization:

The clustering process is methodically carried out employing both K-Means clustering and hierarchical clustering. Hierarchical clustering is explored with a range of linkage strategies including Single, Complete, Average, Centroid, and Ward, and utilizes distinct distance metrics such as Euclidean, Cityblock, Cosine, Correlation, and Chebyshev. To illustrate this comprehensive analysis, Table 2 presents a tabulated summary of the Hierarchical clustering linkages and corresponding distance metrics harnessed within the clustering process.

Table 2. Hierarchical Clustering Linkages and Distance Metrics employed

Clustering Linkage	**Distance Metrics Employed**
Single	Euclidean, Cityblock, Cosine, Correlation, Chebyshev
Complete	Euclidean, Cityblock, Cosine, Correlation, Chebyshev
Average	Euclidean, Cityblock, Cosine, Correlation, Chebyshev
Centroid	Euclidean
Ward	Euclidean

The resultant dendrograms, generated for each combination of linkage and distance metric, are meticulously scrutinized. These visualizations serve as a dynamic tool for identifying effective

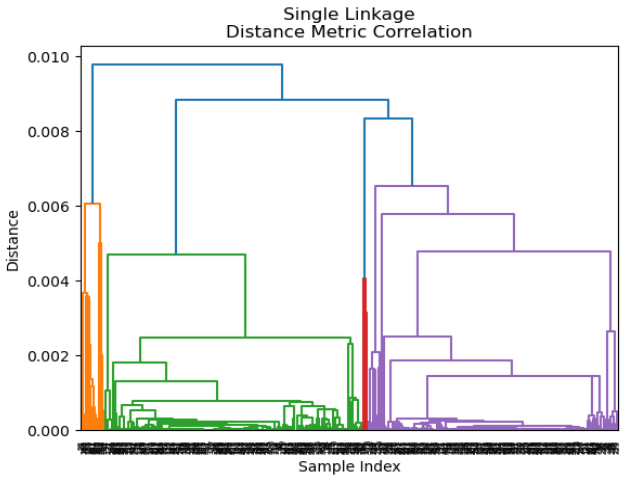

Fig. 3. Dendrogram of Single linkage clusters

clustering approaches based on the cohesiveness of formed

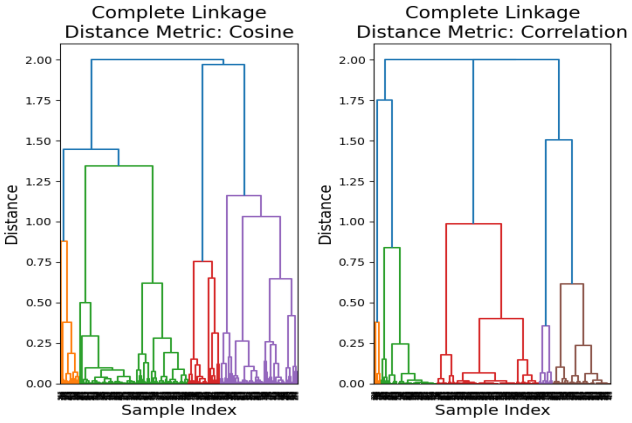

Fig. 4. Dendrograms of Complete linkage clusters

clusters within the dendrogram structure and their hierarchical levels. Figures 3, 4, and 5 offer visual insights into cluster formations employing Single, Complete, and Average linkages with various distance metrics, while Figure 6 provides a snapshot of clustering outcomes using Centroid and Ward linkages combined with Euclidean distance metrics.

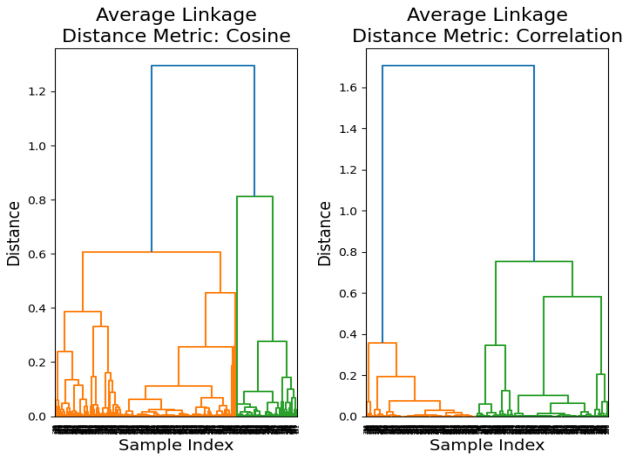

Fig. 5. Dendrograms of Average linkage clusters

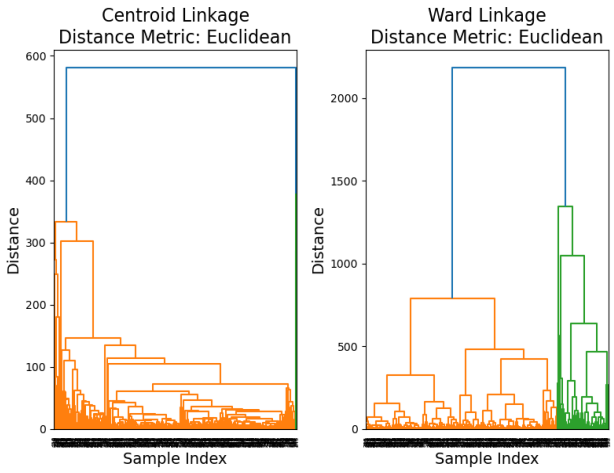

Fig. 6. Dendrograms of Centroid & Ward linkage clusters

The study delves into the intricate dendrogram visualizations, unearthing the underlying clustering structures with exceptional clarity. These sophisticated methodologies adeptly partition companies based on their financial metrics, and the insightful dendrogram visualizations significantly enhances comprehension of the intricate interconnections binding the chosen undervalued companies. The tabulated data in Table 3 provides an overview of the distribution of instances within each cluster.

Table 3. Number of Instances in Each Cluster

Clustering Method	Cluster 1	Cluster 2	Cluster 3
K-Means (K=3)	386	81	3
Single Linkage - Correlation	224	225	21
Complete Linkage - Cosine	254	153	63
Complete Linkage - Correlation	141	125	204
Average Linkage - Cosine	135	118	217
Average Linkage - Correlation	74	184	212
Ward - Euclidean	371	95	4

Thorough examination of the dendrograms reveals profound insights. As illustrated in Table 3, the diversity exhibited by clusters resulting from various linkage and distance metric combinations is compelling. The selection of distinct branches across different clusters is evident within these combinations:

1. Single Linkage – Correlation
2. Complete Linkage – Cosine
3. Complete Linkage – Correlation
4. Average Linkage – Cosine
5. Average Linkage – Correlation

The clustering process is rooted in annual performance, employing the mode of the cluster number for classification. Consequently, a company is assigned to a cluster based on the predominant cluster it aligns with over the past decade. This approach ensures a robust and reliable company-to-cluster association, facilitating a comprehensive understanding of the clustering outcome.

3.5 Analysis of Clustered Data and Model Comparison:

In this phase of the study, an exhaustive examination of the clustered data is undertaken, complemented by a rigorous comparative analysis of the diverse clustering methodologies employed. The tabulated data, as elucidated in Table 4, provides a succinct overview of the allocation of selected companies across individual clusters. Notably, certain companies display bi-modal characteristics, resulting in their simultaneous assignment to two distinct clusters, thus contributing to an aggregate count exceeding 50 companies. It is worth highlighting that, in contrast, the K-Means clustering method and Ward clustering utilizing the Euclidean distance metric exhibit limited diversity within their respective clusters. Consequently, these particular clusters have been omitted from subsequent analyses for their restricted discriminatory capacity.

Table 4. Number of Companies in Each Cluster

Clustering Method	Cluster 1	Cluster 2	Cluster 3
Average Linkage - Correlation	9	20	28
Average Linkage - Cosine	12	8	34
Complete Linkage - Correlation	19	14	25
Complete Linkage - Cosine	37	11	7
Single Linkage - Correlation	30	27	1
Ward - Euclidean	49	2	1
K-Means (K=3)	7	1	48

4. Results

The study assesses the influence of identified clusters of undervalued companies on trading activities and returns, following the approach outlined previously. It provides a comparative analysis of total returns and average return per trade for the selected undervalued companies in contrast to the benchmark Nifty50 companies.

4.1 Clustered Data Insights for Total Returns (%) of clusters:

The analysis of clustered data for a shorter time horizon furnishes valuable insights into the performance of different clusters based on selected technical indicators. As illustrated in Figure 7, a box plot illustrates the distribution of total returns (%) across various clusters for five specific technical indicators over a 365-day period. Figure 8 extends this analysis over a 10-year horizon, showcasing the distribution of total returns (%) across the same clusters and technical indicators. To comprehensively evaluate cluster performance, five combinations of technical indicators were employed to compare returns among the 50 undervalued companies and Nifty50 companies. This facilitates an assessment of cluster performance relative to both the entire cohort of 50 companies and the benchmark Nifty50 companies' returns.

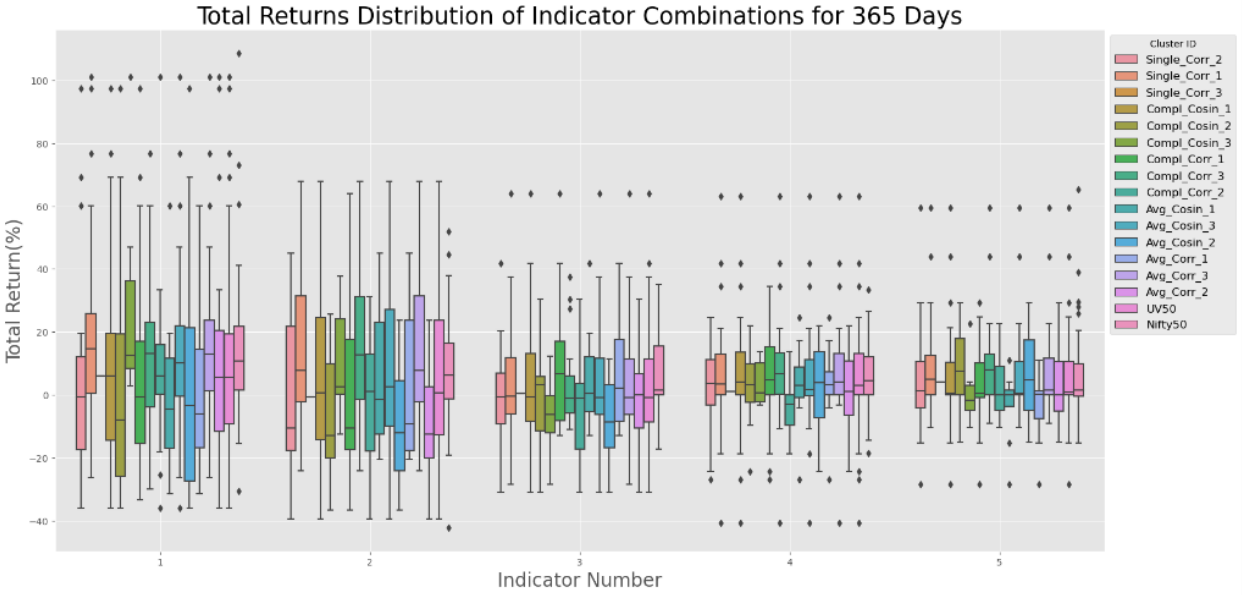

Fig. 7. Summary of Total Returns Distribution for 365 Days

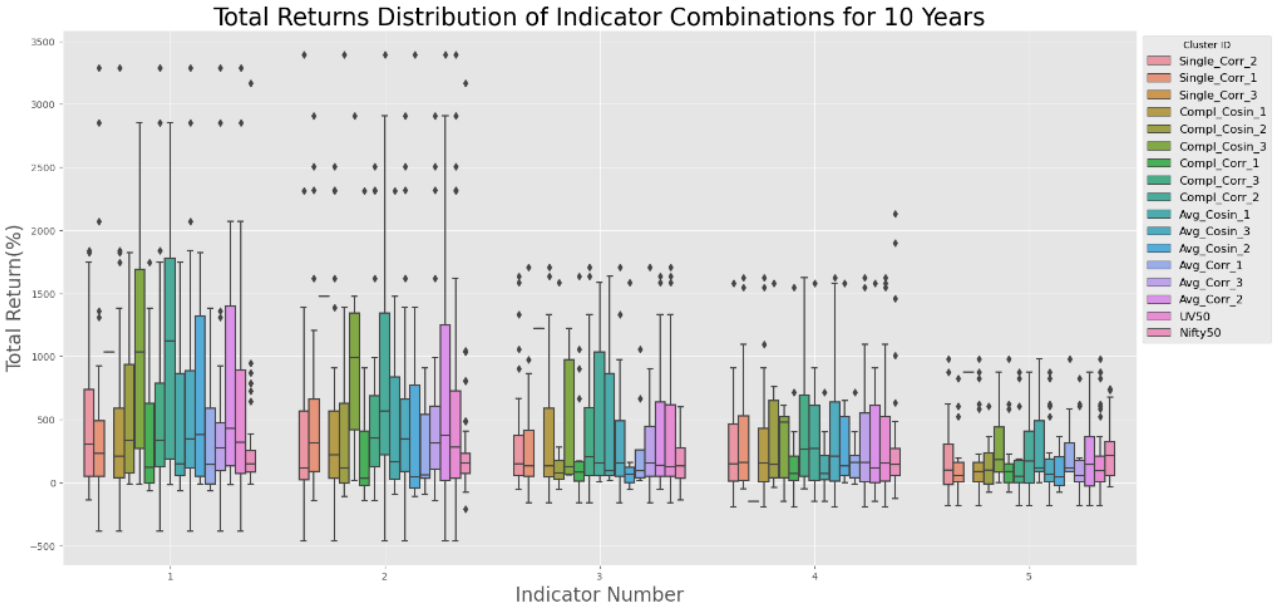

Fig. 8. Summary of Total Returns Distribution for 10 years

List of indicators applied for stock entry are listed below in Table 5. Table 6 offers a sample numerical summary of the total returns (%) for 10 years for an Indicator combination distribution within each cluster for technical indicator number 3.This table encompasses essential statistics, including data point count, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values. These statistics provide a holistic view of total returns of variations across different clusters and technical indicators, contributing to further analysis and decision-making.

Table 5. List of selective Entry and Exit Conditions

S. No.	Entry Condition		
	CCI	OBV	SMA
1	Uptrend	Neutral	NA
2	Uptrend	Downtrend	NA
3	Uptrend	Downtrend	Downtrend
4	Uptrend	Downtrend	Uptrend
5	Downtrend	Neutral	Uptrend
Exit Condition: 10% Trailing Stop Loss			

Table 6. Distribution of Total Returns (%) for 10 Years of an Indicator Combination

Clustering Method	Cluster ID	Cluster Size	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
Average-Correlation	Cluster 1	9	17.6	38.0	89.7	256.0	1054.9
	Cluster 2	20	-58.0	49.9	138.1	635.7	1637.01
	Cluster 3	28	-159.3	51.1	167.6	513.1	2263.26
Average-Cosine	Cluster 1	12	17.6	55.5	91.1	861.4	1637.01
	Cluster 2	8	-58.0	-3.9	62.6	117.4	1587.39
	Cluster 3	34	-159.3	50.0	167.6	564.6	2263.26
Complete-Correlation	Cluster 1	19	-159.3	8.6	82.9	165.4	1637.01
	Cluster 2	14	5.5	57.2	199.9	1223.3	2263.26
	Cluster 3	25	-159.3	52.1	212.4	675.3	2263.26
Complete-Cosine	Cluster 1	37	-159.3	41.4	128.3	590.4	1707.54
	Cluster 2	11	-58.0	24.2	77.4	175.9	1587.39
	Cluster 3	7	57.2	81.3	548.9	1160.9	2263.26
Single-Correlation	Cluster 1	30	-159.3	50.0	138.1	461.6	2263.26
	Cluster 2	27	-58.0	55.1	148.5	372.4	1637.01
	Cluster 3	1	1223.3	1223.3	1223.3	1223.3	1223.26
Undervalued 50		50	-159.3	46.8	131.1	669.8	2263.26
Nifty50		50	-138.6	29.3	128.5	276.4	597.38

4.2 Clustered Data Insights for Average Returns (%) Per Trade:

Figure 9 highlights the distribution of average returns (%) per trade across different clusters and specific technical indicators over a 365-day period. Table 6 complements Figure 9 by presenting comprehensive numerical data for the distribution of average returns per trade within each cluster for the selected technical indicators. These statistics offer valuable insights into how average returns per trade vary across clusters and technical indicators, assisting in decision support. The analysis of clustered data over both short and long-term horizons provides a holistic view of the performance of different clusters, aiding researchers and practitioners in portfolio selection and financial decision support.

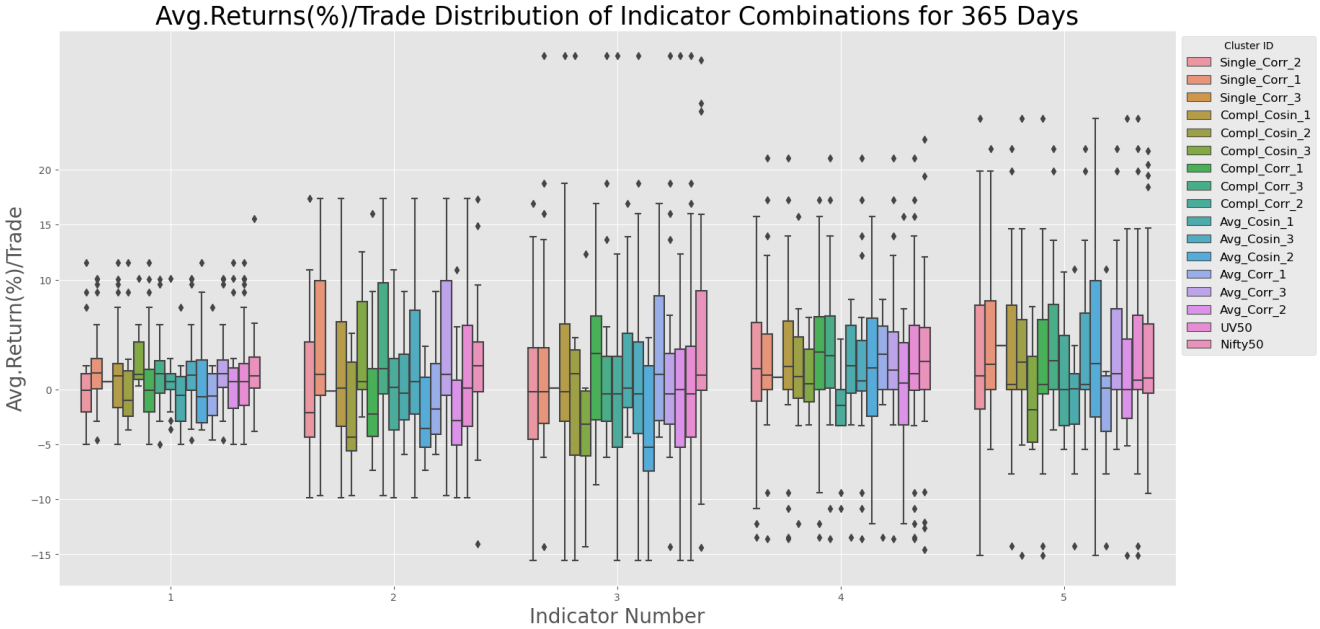

Fig. 9. Average Returns (%)/ Trade Distribution for 365 Days

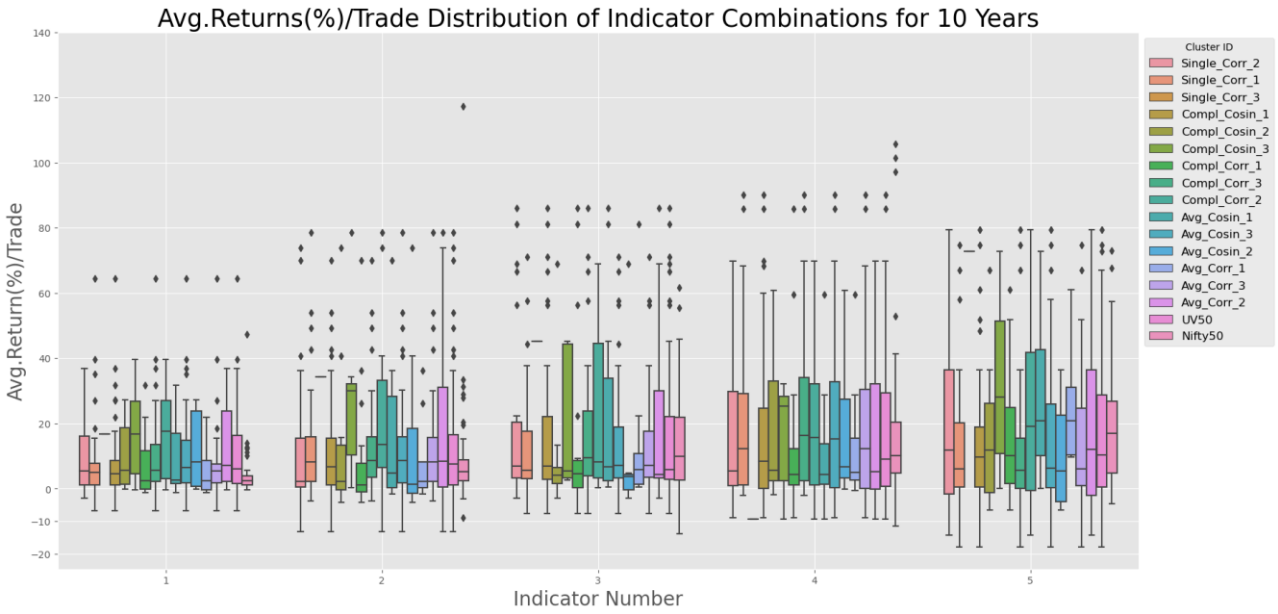

Fig. 10. Average Returns (%)/ Trade Distribution for 10 Years

Table 7. Distribution of Average Returns (%)/ Trade for 10 Years of an Indicator Combination

Clustering Method	Cluster ID	Cluster Size	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
Average-Correlation	Cluster 1	9	-0.55	2.60	5.11	15.41	59.62
	Cluster 2	20	-9.22	-0.19	5.33	32.08	69.81
	Cluster 3	28	-8.94	0.09	12.24	30.46	90.13
Average-Cosine	Cluster 1	12	-9.22	1.37	4.42	13.83	59.62
	Cluster 2	8	-0.19	3.21	6.69	27.53	60.83
	Cluster 3	34	-8.94	0.40	15.31	32.85	90.13
Complete-Correlation	Cluster 1	19	-8.94	1.17	4.42	12.24	85.90
	Cluster 2	14	-9.22	1.17	15.68	32.08	69.81
	Cluster 3	25	-2.06	2.54	16.29	34.15	90.13
Complete-Cosine	Cluster 1	37	-8.94	0.09	8.53	24.62	90.13
	Cluster 2	11	-1.89	2.50	5.73	33.11	60.83
	Cluster 3	7	-9.22	2.36	25.33	28.25	32.08
Single-Correlation	Cluster 1	30	-2.06	1.21	12.24	29.09	90.13
	Cluster 2	27	-8.94	0.86	5.53	29.77	69.81
	Cluster 3	1	-9.22	-9.22	-9.22	-9.22	-9.22
Undervalued 50		50	-9.22	0.66	9.00	29.29	90.13
Nifty50		50	-11.45	4.79	10.24	20.31	105.63

5. Conclusions:

The statistical significance of the clustering results and model performances was rigorously evaluated, shedding light on their practical implications. The clustering analysis conducted in this study yielded significant insights into the performance of different clustering methods, providing valuable implications for portfolio selection and financial decision support systems. Several key findings and conclusions emerge from the research:

5.1 Efficiency of Clustering Methods:

The efficiency of various clustering methods was assessed, revealing that K-Means clustering and Ward linkage in Hierarchical clustering did not produce diverse clusters. In contrast, Hierarchical clustering methods using Single-Correlation, Average-Correlation, Average-Cosine, Complete-Correlation, and Complete-Cosine linkages demonstrated greater diversity, enhancing their suitability for creating distinct stock clusters.

5.2 Performance Relative to Benchmarks:

To evaluate cluster performance, comparisons were made between the clusters, the entire cohort of 50 undervalued companies, and the benchmark Nifty50 companies' returns. Remarkably, certain clusters outperformed both the cohort of 50 undervalued companies and the benchmark Nifty50 in terms of returns, showcasing their potential for investment.

5.3 Optimal Clusters:

Based on a holistic assessment, which considers factors such as fewer negative returns and high returns at the 50th and 75th percentiles, several clusters have emerged as optimal within their respective hierarchical clustering methods. These include Single-Correlation-Cluster 2, Complete-Cosine-Cluster 3, Complete-Correlation Cluster 2, Average-Cosine Cluster 2, and Average-Correlation Cluster 2.

Based on overall performance considerations, specific clusters emerged as optimal choices within their respective hierarchical clustering methods. Notably, clusters such as Single-Correlation-Cluster 2, Complete-Cosine-Cluster 3, Complete-Correlation Cluster 2, Average-Cosine Cluster 2, and Average-Correlation Cluster 2 displayed fewer negative returns, higher returns at the 50th and 75th percentile, and robust performance compared to other clusters and benchmarks.

Notably, three clusters, namely Complete-Cosine-Cluster 3, Complete-Correlation Cluster 2, and Average-Correlation Cluster 2, consistently outperformed other cluster techniques in terms of returns. These clusters have demonstrated superior performance compared to both the entire cohort of 50 undervalued companies and the benchmark Nifty50 companies' returns.

The clustering analysis conducted in this research contributes valuable insights to the field of portfolio selection and financial decision support. It offers a methodological framework for investors to identify undervalued stocks with the potential for superior returns. By leveraging advanced clustering techniques, this approach empowers investors to make data-driven decisions and enhance their financial portfolios.

References

[1] J. Li, X. Wang, S. Ahmad, X. Huang, and Yousaf Ali Khan, "Optimization of investment strategies through machine learning.," Heliyon, vol. 9, no. 5, pp. e16155–e16155, May 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e16155.

[2] M. Leippold, Q. Wang, and W. Zhou, "Machine-Learning in the Chinese Factor Zoo," SSRN Electronic Journal, 2020, doi: https://doi.org/10.2139/ssrn.3754339.

[3] V. Azevedo and C. Hoegner, "Enhancing stock market anomalies with machine learning," Review of Quantitative Finance and Accounting, vol. 60, no. 8, Aug. 2022, doi: https://doi.org/10.1007/s11156-022-01099-z.

[4] K. Ray, "Artificial Intelligence and Value Investing," The Journal of Investing, vol. 27, no. 1, pp. 21–30, Feb. 2018, doi: https://doi.org/10.3905/joi.2018.27.1.021.

[5] R. Chopra and G. D. Sharma, "Application of Artificial Intelligence in Stock Market Forecasting: A Critique, Review, and Research Agenda," Journal of Risk and Financial Management, vol. 14, no. 11, p. 526, Nov. 2021, doi: https://doi.org/10.3390/jrfm14110526.

[6] L. Chen, M. Pelger, and J. Zhu, "Deep Learning in Asset Pricing," Management Science, vol. 12, no. 10, Feb. 2023, doi: https://doi.org/10.1287/mnsc.2023.4695.

[7] F. Ren, Y.-N. Lu, S.-P. Li, X.-F. Jiang, L.-X. Zhong, and T. Qiu, "Dynamic Portfolio Strategy Using Clustering Approach," PLOS ONE, vol. 12, no. 1, p. e0169299, Jan. 2017, doi: https://doi.org/10.1371/journal.pone.0169299.

[8] T. Liu, "Non-Negative Matrix Factorization for Stock Market Pricing," International Conference on Engineering and Informatics, Jan. 2009, doi: https://doi.org/10.1109/bmei.2009.5304773.

[9] K. Marvin and S. Bhatt, "Creating Diversified Portfolios Using Cluster Analysis," 2015. Available: https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf

[10] M. Nourahmadi and H. Sadeqi, "Portfolio Diversification Based on Clustering Analysis," Iranian Journal of Accounting, Auditing and Finance, vol. 7, no. 3, pp. 1–16, Aug. 2023, doi: https://doi.org/10.22067/ijaaf.2023.43078.1092.

[11] J. Kim, Yae Jean Kim, S. Jung, Jong Ho Moon, and J. Kwon, "The effects of cluster collaboration and the utilization of big data on business performance: A research based on the expansion of open innovation and social capital," African Journal of Science, Technology, Innovation and Development, vol. 14, no. 4, pp. 1032–1049, Jun. 2021, doi: https://doi.org/10.1080/20421338.2021.1925394.

[12] Ahmad Zaman Khan and Mukesh Kumar Mehlawat, "Dynamic portfolio optimization using technical analysis-based clustering," International Journal of Intelligent Systems, vol. 37, no. 10, pp. 6978–7057, Mar. 2022, doi: https://doi.org/10.1002/int.22870.

[13] S. R. Nanda, B. Mahanty, and M. K. Tiwari, "Clustering Indian stock market data for portfolio management," Expert Systems with Applications, vol. 37, no. 12, pp. 8793–8798, Dec. 2010, doi: https://doi.org/10.1016/j.eswa.2010.06.026.

[14] A. Cirulli, M. Kobak, and U. Ulrych, "Portfolio Construction with Hierarchical Momentum," SSRN Electronic Journal, vol. 6, no. 1, 2022, doi: https://doi.org/10.2139/ssrn.4125072.

[15] S. Roy, R. Bhattacharya, F. Advisor, and H. Khajar Bashir, "Identifying Homogeneity of Small-Cap Stocks in Indian Market: A Data Mining Approach," Apr. 2019.

[16] G. S. KONSTANTINOV and F. J. FABOZZI, "PORTFOLIO VOLATILITY SPILLOVER," International Journal of Theoretical and Applied Finance, vol. 25, no. 04n05, Jun. 2022, doi: https://doi.org/10.1142/s0219024922500194.

[17] Y. Wang and T. Aste, "Dynamic portfolio optimization with inverse covariance clustering," Expert Systems With Applications, vol. 213, no. 1, pp. 118739–118739, Mar. 2023, doi: https://doi.org/10.1016/j.eswa.2022.118739.

[18] E. J. Menvouta, S. Serneels, and T. Verdonck, "Portfolio optimization using cellwise robust association measures and clustering methods with application to highly volatile markets," The Journal of Finance and Data Science, vol. 9, no. 6, p. 100097, Nov. 2023, doi: https://doi.org/10.1016/j.jfds.2023.100097.

[19] T. Lim and C. Sin Ong, "Portfolio Diversification Using Shape-Based Clustering," The Journal of Financial Data Science, vol. 3, no. 1, pp. 111–126, Dec. 2020, doi: https://doi.org/10.3905/jfds.2020.1.054.

[20] O. Alqaryouti, Tarek Farouk, and Nur Siyam, "Clustering Stock Markets for Balanced Portfolio Construction," Advanced Intelligent Systems and Informatics, vol. 53, no. 1, pp. 577–587, Sep. 2018, doi: https://doi.org/10.1007/978-3-319-99010-1_53.

[21] S. Goudarzi, M. Jafari, and A. Afsar, "International Journal of Economics and Financial Issues A Hybrid Model for Portfolio Optimization Based on Stock Clustering and Different Investment Strategies," International Journal of Economics and Financial Issues, vol. 7, no. 3, pp. 602–608, 2017.

[22] S. Roy and R. Bhattacharya, “Clustering Mid-Cap Stocks in Indian Market using Multi-Variate Data Analysis Technique,” Indian Journal of Economics and Development, vol. 7, no. 6, pp. 1–10, Jun. 2019.

[23] S. Burney, H. Tariq, and T. Jilani, “A Portfolio Optimization Algorithm Using Fuzzy Granularity Based Clustering AND BIOTECHNOLOGY View project ICT e-Helath View project,” 2019.

[24] K. S. Pritam, T. Mathur, S. Agarwal, S. K. Paul, and A. Mulla, “A novel methodology for perception-based portfolio management,” Annals of Operations Research, vol. 315, no. 2, pp. 1107–1133, Feb. 2022, doi: https://doi.org/10.1007/s10479-022-04530-9.

[25] A. M. Harsha and V. V. S. K. Rao, “Addressing Challenges in Stock Selection: A Financial Decision Support System Approach,” Journal of Research Administration, vol. 5, no. 2, pp. 4497–4510, Nov. 2023, [Online]. Available: https://journalra.org/index.php/jra/article/view/602