“Expanding portfolio diversification through cluster analysis beyond traditional volatility”

AUTHORS	Mykhailo Kuzheliev Dmytro Zherlitsyn Ihor Rekunenko Alina Nechyporenko Sergii Stabias

ARTICLE INFO	Mykhailo Kuzheliev, Dmytro Zherlitsyn, Ihor Rekunenko, Alina Nechyporenko and Sergii Stabias (2025). Expanding portfolio diversification through cluster analysis beyond traditional volatility. *Investment Management and Financial Innovations*, *22*(1), 147-159. doi:10.21511/imfi.22(1).2025.12
DOI	http://dx.doi.org/10.21511/imfi.22(1).2025.12
RELEASED ON	Thursday, 23 January 2025
RECEIVED ON	Monday, 09 December 2024
ACCEPTED ON	Wednesday, 15 January 2025
LICENSE	This work is licensed under a Creative Commons Attribution 4.0 International License
JOURNAL	"Investment Management and Financial Innovations"
ISSN PRINT	1810-4967
ISSN ONLINE	1812-9358
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
30	3	3

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine

www.businessperspectives.org

Mykhailo Kuzheliev (Ukraine), **Dmytro Zherlitsyn** (Bulgaria), **Ihor Rekunenko** (Ukraine), **Alina Nechyporenko** (Ukraine), **Sergii Stabias** (Ukraine)

EXPANDING PORTFOLIO DIVERSIFICATION THROUGH CLUSTER ANALYSIS BEYOND TRADITIONAL VOLATILITY

Abstract

The study reviews the application of machine learning tools in financial investment portfolio management, focusing on cluster analysis for asset allocation, diversification, and risk optimization. The paper aims to explore the use of clustering analysis to broaden the concept of portfolio diversification beyond traditional volatility metrics. An open dataset from Yahoo Finance includes a ten-year historical period (2014–2024) of 130 actively traded securities from international stock markets used. Dataset selection prioritizes top liquidity and trading activity. Python analytical tools were employed to clean, process, and analyze the data. The methodology combines classical Markowitz optimization with clustering analysis techniques, highlighting variance-return trade-offs. Various asset characteristics, including annualized return, standard deviation, Sharpe ratio, correlation with indices, skewness, and kurtosis, were incorporated into the clustering models to reveal hidden patterns and groupings among financial assets. Results show that while clustering enhances insights into asset diversity, classical approaches remain historically superior in optimizing risk-adjusted returns. This study concludes that clustering complements, rather than replaces, classical methods by broadening the understanding of diversification and addressing many diversity factors, such as metrics of the technical, graphical, and fundamental analysis. The paper also introduces the diversity rate based on clustering, which measures the variance balance by all features within and between clusters, providing a broader perspective on diversification beyond traditional metrics. Future research should investigate dynamic clustering techniques, integrate fundamental economic indicators, and develop adaptive models for effective portfolio management in evolving financial markets.

Keywords financial investment, asset allocation, financial security, portfolio, diversification, machine learning (ML), cluster analysis, financial market, Python analytical tools

JEL Classification C63, C61, D53, G11, G17

Received on: 9th of December, 2024
Accepted on: 15th of January, 2025
Published on: 23rd of January, 2025

Mykhailo Kuzheliev, Doctor of Economics, Professor, Professor of the Department of Finance, National University of "Kyiv-Mohyla Academy", Ukraine.

Dmytro Zherlitsyn, Dr.Sc. (habil.) in Economics, Professor, Researcher of the Institute of Entrepreneurship, University of National and World Economy, Bulgaria.

Ihor Rekunenko, Doctor of Economics, Professor, Head of the Oleg Balatskyi Department of Management, Sumy State University, Ukraine. (Corresponding author)

Alina Nechyporenko, Ph.D. in Economics, Associate Professor of the Department of Finance, Borys Grinchenko Kyiv Metropolitan University, Ukraine.

Sergii Stabias, Ph.D. in Economics, Lecturer of the Department of Economic and Sociological Disciplines, Private institution of professional preliminary education "College of Information Technologies "STEP", Ukraine.

Conflict of interest statement:
Author(s) reported no conflict of interest

INTRODUCTION

Financial markets underwent a dramatic transformation in 2008, revealing unprecedented volatility and exposing the limitations of traditional portfolio management methods. The global financial crisis led to significant losses for investors and underscored the inadequacy of classical tools in addressing complex, dynamic market risks. For instance, as measured by the VIX index, market volatility peaked at a record high of over 89 points in October 2008, a stark indicator of instability (Yahoo Finance, 2024). However, traditional methods did not account for the level of volatility, requiring investors to rethink risk management and explore advanced strategies for portfolio diversification.

Current research on innovative methods and techniques has expanded significantly. It has emphasized the constraints of traditional approaches in addressing multidimensional risk factors, particularly in high-frequency trading and behavioral biases. A bibliometric study

analyzing over 589 articles published between 2003 and 2023 reveals that approximately 71% of research in financial technical analysis now focuses on integrating machine learning (ML) techniques, such as sentiment analysis and algorithmic trading (Gallastegui et al., 2024; Inani et al., 2024; Sang, 2024). Therefore, scientists, financial specialists, and investors actively use ML and Big Data techniques. Thus, the problem of evaluating the accuracy and efficiency of the different time series models to use the virtual financial vehicle as a competitive investment asset is urgent for investors and data scientists.

Thus, the need to redefine diversification stems from the complexity of contemporary financial markets, where static, one-dimensional metrics and traditional optimization methods are insufficient for capturing dynamic interdependencies among assets.

1. LITERATURE REVIEW AND HYPOTHESES

The evolving complexity of financial markets has highlighted limitations in traditional portfolio optimization methods, necessitating the integration of advanced analytical tools such as machine learning, e.g., cluster analysis. Many studies and scientific publications focus on different problems of technical and fundamental indicators of the dynamic of financial market assets and the practical application of various types of time series, machine learning, and artificial intelligence models for predicting price trends.

The current research landscape is rich with studies exploring the interplay between ML techniques and portfolio management, particularly in forecasting and risk assessment. This landscape is characterized by the convergence of classical financial theories with cutting-edge ML tools, which has led to novel insights into financial markets. Researchers have focused on addressing the limitations of traditional models, such as the Markowitz framework, by introducing methods that incorporate multidimensional asset characteristics, behavioral patterns, and advanced statistical metrics.

Much of the current research landscape explores ML's role in **enhancing forecasting capabilities**. ML's utility in handling complex, high-risk environments is further highlighted by Fantazzini and Zimin (2020), who introduce advanced models like GARCH and the Zero Price Probability model for cryptocurrency market and credit risk assessments. Derbentsev et al. (2021) reinforce this with ensemble-based ML approaches, demonstrating superior accuracy and robustness in cryptocurrency price forecasting. Similarly, Aguirre et al. (2020) show the effectiveness of Genetic Algorithms in optimizing technical indicators and providing an adaptable and robust solution for trading strategies. Zmuk and Jošić (2020) describe ML methods such as linear regression, Gaussian processes, and Neural Networks to predict stock market indices with high precision, particularly for shorter time horizons. Liew and Mayster (2018) extend ML applications to ETF predictions, demonstrating the superior accuracy of ML algorithms in capturing complex relations. Korstanje (2021) uses advanced time-series forecasting techniques (Prophet, LSTMs, DeepAR, etc.) in complex financial data prediction. These contributions underscore ML's utility in improving prediction accuracy and market analysis. Apalkova et al. (2022) examine various aspects of how price levels and purchasing power influence environmental performance across different countries, utilizing the machine learning capabilities of RapidMiner.

The second pillar of the current research landscape is on **ML-based portfolio optimization**, revealing its transformative potential in redefining diversification strategies. Clarissa and Koesrindartoto (2024) propose a dynamic portfolio optimization strategy that outperforms traditional benchmarks using predictive models and adaptive optimization. Heaton et al. (2017) apply deep learning models to detect non-linear patterns in complex financial data and determine the way for widespread adoption in portfolio optimization. Feng et al. (2024) extend these approaches to sustainable investments, showing how ESG and SDG sentiment analysis improves portfolio performance and aligns investments with long-term societal goals. Bhama (2024) focuses on equity markets, revealing how moving average strategies in high-volatility portfolios outperform traditional Buy-and-Hold approaches. López de Prado (2016) introduces clustering-based ML approaches to improve diversification and outperform traditional benchmarks in portfolio

http://dx.doi.org/10.21511/imfi.22(1).2025.12

optimization. Owen (2023), Jain P. and Jain S. (2012) employ hierarchical clustering to enhance portfolio performance by structuring assets based on correlation and risk parity strategies against covariance estimation errors. Pinelis and Ruppert (2021) advance portfolio allocation strategies by integrating return- and volatility-timing using Random Forest models, achieving substantial improvements in Sharpe ratios and maximum drawdowns. Leung et al. (2023) developed a machine learning-based portfolio recommendation system incorporating big data analytics for personalized investment strategies.

ML tools also play a critical role in identifying **market anomalies** and **behavioral patterns**. Viebig (2020) applies Support Vector Machines to detect irrational exuberance in financial markets, predicting subsequent abnormally low returns and helping investors mitigate risks. Aiche et al. (2024) explore the application of artificial intelligence in constructing and managing cybersecurity stock portfolios. Using advanced machine learning techniques (Random Forests, Support Vector Machines) and natural language processing for sentiment analysis, the study combines predictive modelling with mean-variance optimization to deliver superior portfolio performance. Aziz et al. (2021) review ML applications in finance, emphasizing the increasing importance of sentiment analysis and text-based ML methods in providing novel insights or behavioral patterns in financial data.

Finally, part of the current research landscape of ML and clustering tools in **enhancing decision-making processes** is further underscored in several investigations. Babenko et al. (2021) review classical ML methods like regression and clustering, demonstrating their adaptability across various levels of economic analysis. Mints (2017) categorizes data mining tasks in finance, highlighting ML's potential to automate complex processes and improve forecasting accuracy. Kuzheliev et al. (2019, 2020) use traditional econometric analysis tools to explore and predict Ukraine's macroeconomic and financial indicators. Glazunova et al. (2021) focus on improvements in education and professional skills and propose structured approaches to enhance digital intelligence in economists, including project-based learning and real-world applications. The investigations have confirmed the efficiency of ML tools in improving financial literacy and decision-making results.

Collectively, these studies highlight the growing integration of machine learning, e.g. clustering analysis, in financial research. They demonstrate the versatility of machine learning tools in market forecasting, portfolio management, risk assessment, and a wide range of problems.

Although significant progress has been made in applying machine learning to financial portfolio management, essential gaps still need to be identified. Most of the existing research focuses on the technical aspects of applying ML tools and IT technologies. Fundamental economic factors remain outside the scope of the study. Although some researchers (Owen, 2023; Jain & Jain, 2019) raise the issue of enhanced portfolio diversification, the traditional concept of optimization based on the apparent relationship between volatility and average return (Markowitz, 1952) remains the same. Consequently, it is essential to explore applying specific ML methods in managing financial investment portfolios, clarify the definition of diversification, and establish key diversification metrics. By addressing these issues, clustering algorithms and tools can become even more effective in supporting dynamic investment decisions based on data and expanding economic theory concepts.

The paper aims to explore the use of clustering analysis to broaden the concept of portfolio diversification beyond traditional volatility metrics. Specifically, it examines how clustering techniques can uncover hidden patterns and groupings among financial assets. It offers insights into more effective diversification strategies that include multidimensional asset characteristics, such as behavioral patterns, correlations, and advanced statistical metrics.

Based on the paper's aim, the key research hypotheses are as follows

H1: Cluster analysis algorithms consistently outperform classical statistical methods in constructing portfolios with the highest rate of returns or optimal risk-return scenarios.

H2: Diversification of a financial investment portfolio can be measured not only by volatility indicators as a variation of returns.

2. METHODOLOGY AND DATA

The study is based on international stock market asset prices, which are analyzed using Python's advanced analytical tools. This approach comprehensively evaluates market dynamics and asset behaviour through robust and widely adopted computational methods.

The study uses open data from Yahoo Finance (Yahoo, 2024). This is a trusted source of financial information. The selection of assets is guided by their trading activity and relevance in financial markets, leveraging tools such as the Yahoo Finance screener API (Zherlitsyn, 2024). The screener function of the yahooquery[1] Python library identifies assets' tickers based on trading volume and liquidity criteria. The data include critical financial attributes such as historical prices, daily returns, and associated key indicators, forming the basis for the analytical processes in this study. This approach ensures that the selected assets are among the most representative of active trading behavior, providing the most variance dataset for analysis. These asset data are systematically cleaned, filtered, and classified to create statistically homogeneous samples.

The study incorporates a ten-year dataset for stock market analysis (started from 01/01/2014). This period captures trends and significant variations, providing a robust foundation for evaluating asset performance. Extending the time horizon reduces the number of eligible assets (by more than half). Therefore, a year's time series is a practical and optimal choice for including a more significant number of assets. Historical data, including daily prices and their changes, are utilized for financial investment portfolio optimization within this range.

Figure 1 illustrates the study's research steps.

The first step, Data Collection and Cleaning, begins with acquiring financial market data, focusing on a selected list of financial securities using Yahoo Finance API. This involves defining the list of assets based on their trading activity and relevance. The collected data undergo a cleaning process to remove inconsistencies, missing values, and outliers, preparing it for further analysis. Also, it includes calculating percentage changes in daily prices and normalizing the data to create a consistent and analyzable dataset.

Second, Classical Portfolio Optimization is used to build the cleaned data to optimize the portfolio

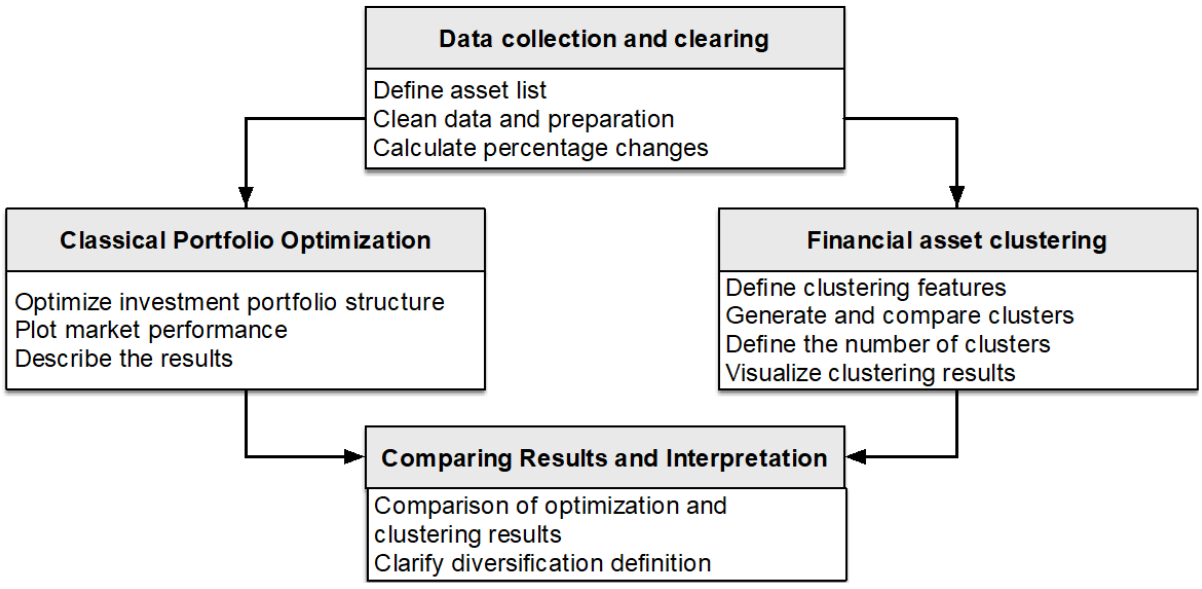

Figure 1. Flowchart of the research

1 https://pypi.org/project/yahooquery/

http://dx.doi.org/10.21511/imfi.22(1).2025.12

structure using the Rate of Return, Variance, and Sharpe ratio (Zherlitsyn, 2024). These optimization metrics are used to identify an optimal allocation of assets that balances risk and return. Results are visualized through a market plot that illustrates the performance of different portfolios and highlights the optimal structure position. This analysis uses historical data for the entire period and Python-based computational tools (Korstaje, 2021) to evaluate and display the findings.

Third, Financial Asset Clustering is used for portfolio optimization. This analyzes financial assets by applying clustering techniques. Based on the cleaned data, features for clustering are defined, and experiments are conducted with different feature sets to understand the impact of diversification. The optimal number of clusters is determined using elbow analysis, ensuring meaningful segmentation. The clustering results are visualized to demonstrate how assets are grouped based on diversity.

Finally, Compare and interpret the results of portfolio optimization and clustering outcomes. This includes checking the Sharpe-ratio-optimal assets within the defined clusters and analyzing the cluster centroids to gain a deeper understanding of their diversity. Clarify the concept of diversification based on clustering.

3. RESULTS

Before applying optimization and clustering techniques, data were gathered and processed. This preparation phase entails selecting, cleansing, and converting raw financial data into a format suitable for analysis. The emphasis is on acquiring comparative and statistically sound information about actively traded assets and their structuring. The study utilized Python's yahooquery library to extract data for the top 250 most active assets across multiple markets using the screener function[2]. This approach ensures a diversified dataset representing the most liquid and actively traded securities.

The raw dataset included historical daily price data for each asset. The initial dataset was incomplete, necessitating a cleaning process and security exceptions for which there is insufficient comparable data. Missing values, duplicates, and anomalies were identified and removed to ensure consistency and reliability. To quantify the daily movement of asset prices, percentage changes were calculated based on adjusted close prices. The prepared dataset was summarized in a structured format. The final dataset includes 130 securities for futures analysis.

The classical method for portfolio optimization relies on the foundational principles established by Harry Markowitz (Markowitz, 1952), incorporating a quantitative focus on balancing variance and return. This study applies the classical criteria (rate of returns and variance) and Sharpe ratio to identify the optimal portfolio composition by maximizing the risk-adjusted return. The Sharpe ratio of portfolio return to its standard deviation evaluates how effectively a portfolio compensates investors for the risks undertaken (Zherlitsyn, 2024).

The classical portfolio optimization results are based on a time horizon from January 1, 2014 to December 1, 2024, and are summarized in Table 1 and Figure 2.

As shown in Table 1, the classical approach to optimizing the structure of an investment portfolio reduces the number of assets to 18. The criterion of maximizing the rate of return leaves only one asset

Table 1. Performance and diversification metrics for optimized portfolios (classical optimization method) from January 1, 2014 to December 1, 2024

Optimization Criteria	Rate of return of the Portfolio	Volatility	Number of Securities in the Portfolio	Diversification Ratio
Maximize Sharpe Ratio	0.38	0.22	18	0.1166
Maximize Rate of Return	0.84	0.83	1	0.063
Minimize Volatility	0.11	0.12	18	0.125

Note: * estimations based on the Yahoo Finance data (Yahoo, 2024).

2 https://yahooquery.dpguthrie.com/guide/screener/

in the portfolio. Initially, the analysis considered data for 130 assets. Examining the Diversification Ratio, which is based on Markowitz principles (Markowitz, 1952) (the ratio of the weighted sum of asset rates of return standard deviations to the portfolio's standard deviation), reveals that the portfolio constructed under the criterion of minimizing volatility achieves the highest diversification. However, this portfolio significantly underperforms in returns compared to the portfolio optimized using the Sharpe ratio. In the latter case, the return increases more than threefold while the Diversification Ratio decreases from 0.125 (for the minimum volatility portfolio) to 0.1166.

Figure 2 visually confirms these results. It depicts various combinations of investment portfolios, and the optimal risk-return ratios, derived using the Monte Carlo simulation method, are highlighted by the colored dots. The plot illustrates the trade-offs between portfolio return (y-axis) and risk, measured as standard deviation. As can be seen, Figure 2 visually supports the data presented in Table 1, demonstrating the distribution of portfolios and the effectiveness of different optimization criteria. The Sharpe-optimal portfolio lies along the efficient frontier, balancing higher returns with moderate risk, while the minimum-volatility and maximum-return portfolios represent the extreme ends of the spectrum.

Thus, this study will use the portfolio structure optimized based on the Sharpe ratio criterion for further comparative analysis. As shown in Table 1, this portfolio includes 18 assets: NVDA, TSLA, MARA, KGC, BTG, AVGO, PANW, IAG, WMT, MSTR, IBN, DXCM, MO, SMCI, NEM, K, CELH, KDP. The statistical characteristics of these assets' rates of return demonstrate the range of asset variability and distribution over the analyzed period. However, diversifying the average rate of return with the level of volatility does not consider other factors and descriptive statistical indicators, such as outliers in returns, which can be regarded as using metrics like the median, quartiles, and similar measures. On the other hand, the selected 18 assets, while statistically significant, do not include some "blue-chip" stocks from the market, such as MSFT or AAPL. Therefore, a cluster analysis method will be applied to the presented dataset to assess the differences between assets further and refine the understanding of diversification.

Clustering analysis was introduced as an advanced method to deepen the understanding of portfolio diversification and capture hidden relationships between assets. Unlike classical approaches that rely on return and volatility metrics, clustering may incorporate a comprehensive set of features derived from classical finan-

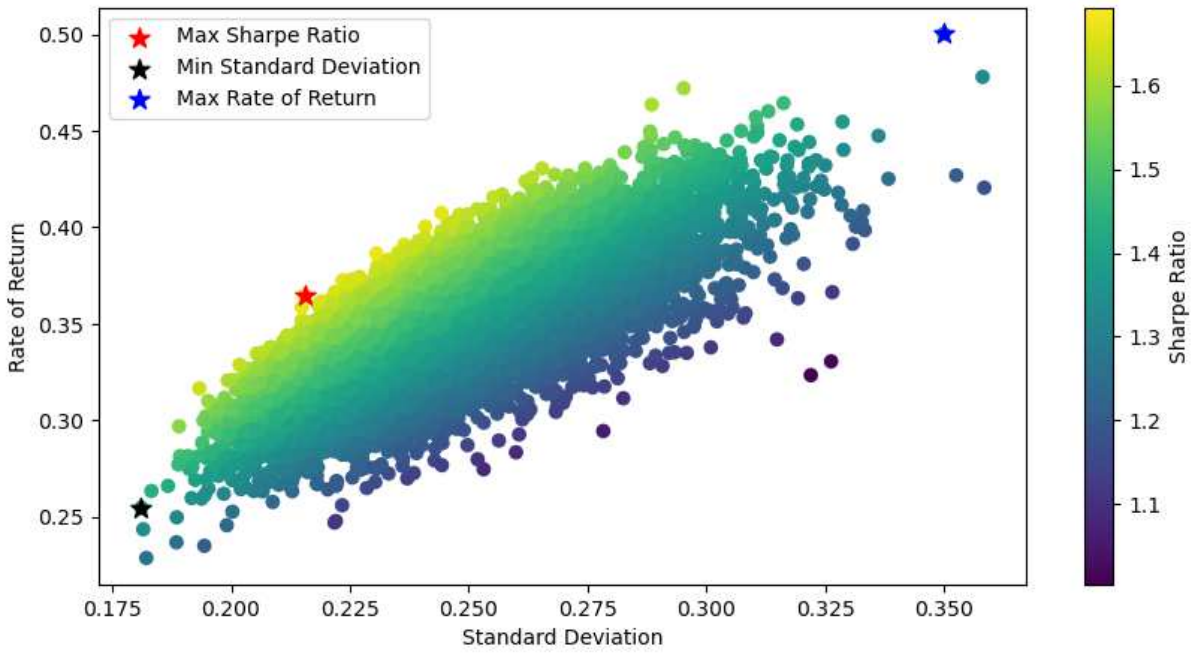

Figure 2. Efficient frontier and simulated portfolio risk-return trade-off

http://dx.doi.org/10.21511/imfi.22(1).2025.12

cial and technical analyses and other statistical indices, like autocorrelation measures within the trading week.

This approach's features lie in capturing assets' multidimensional characteristics. These include their return and risk metrics, behavioral patterns, correlations with broader market indices, and statistical properties such as skewness and kurtosis. These factors provide a holistic view of the assets' performance and potential groupings based on their diversity.

The following features are used for clustering and future investigations:

- Mean, standard deviation of returns, and Sharpe ratio of each asset's average performance and variability like in the classical approach.
- Correlation with the NASDAQ index (^IXIC), indicating the degree to which each asset's returns move in line with the broader market. This is a one-dimensional replacement for the Markowitz correlation matrix.
- Skewness and kurtosis to capture the asymmetry and tail heaviness of the return distributions.
- Autocorrelation of returns across different lags (1 to 5 days) to assess the persistence of patterns in daily returns and detect seasonality within the trading week.
- Moving averages over 21, 63, and 126 trading days, representing short-term, medium-term, and long-term trends.

The feature processing begins with calculating these metrics for each asset in the dataset. For instance, autocorrelation is calculated at multiple lags to explore potential patterns in daily returns, while moving averages provide insights into the asset's price momentum over different time frames. The correlation with the NASDAQ index helps identify how closely each asset aligns with the overall market behaviour.

Asset groups with similar statistical and technical characteristics are identified using all or a combination of these features. The results are summarized in a structured data set, which serves as input for clustering algorithms. Since one of the crucial problems of cluster analysis is determining the optimal number of clusters, at the initial stage, the paper assumes the presence of 18 initial clusters (these results from implementing the principles of the classical Markowitz model based on optimizing the Sharpe ratio). However, in future analyses, all the searching features will be applied for the optimal number of clusters based on the elbow method, Silhouette Score, and visual analysis of the hierarchical cluster analysis dendrogram.

The hierarchical clustering analysis used the features derived from classical and technical analysis and autocorrelation metrics. The feature dataset was pre-processed by scaling all variables to a standard range. This ensured that all features may contribute equally to the clustering process, regardless of their original scales or units. Ward's linkage method was applied for hierarchical clustering. This method minimizes the variance within clusters. The distance metric used was Euclidean distance, which measures the straight-line distance between points in a multidimensional space.

The dendrogram (Figure 3) visually represents the hierarchical relationships among the portfolio's assets. It displays how assets are merged into clusters at different levels of similarity. The color threshold was set at a specific distance level to highlight the most vizuality clusters.

Figure 3 illustrates the most significant differences emerge within 3 to 7 clusters. This number is substantially lower than the assets selected using the classical method (Table 1).

Additional analysis of various asset combinations, based on the elbow method and the Silhouette Score, also revealed that the maximum meaningful differentiation for the examined dataset may be achieved with 17 clusters. This result aligns closely with the outcomes of the classical optimization approach. However, it is worth emphasizing that such a result is contingent on the restricted number of selected features. As the number of factors increases, the differentiation between clusters diminishes. In contrast, clusters in the range of

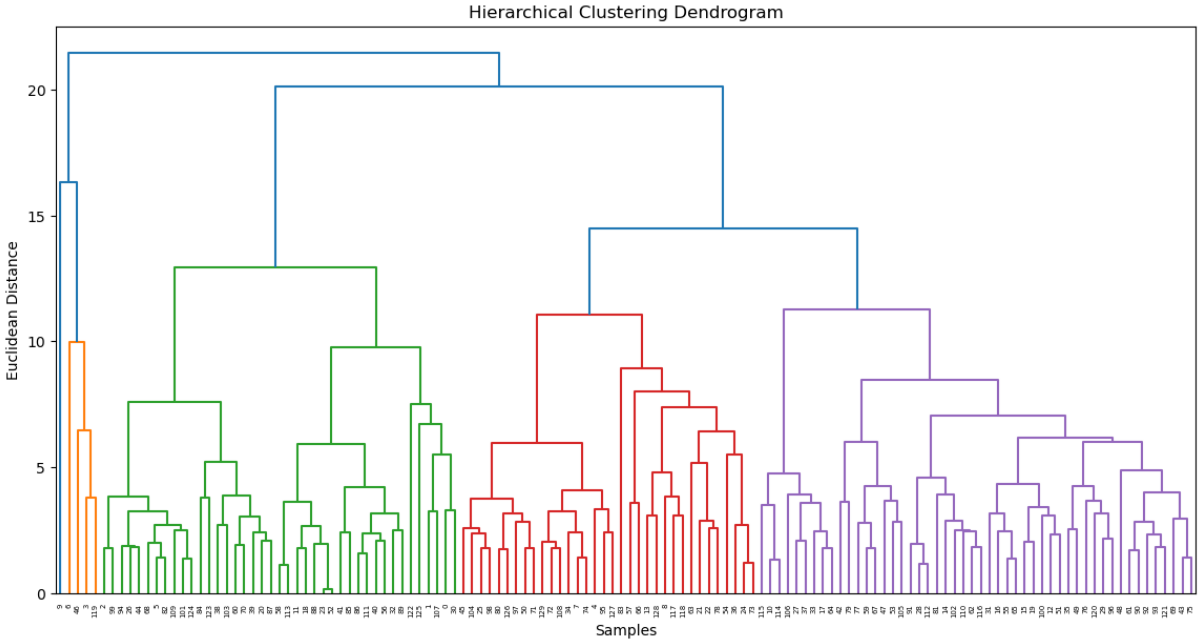

Figure 3. Hierarchical clustering dendrogram for portfolio assets using Ward's method

6 to 7 provide the most stable distinctions across all described features. This stability reflects the data's inherent structure and offers a foundation for further segmentation and portfolio analysis, as shown in Figure 3.

Table 2 presents the recalculated annualized mean returns and standard deviations (volatilities) for each of the six clusters identified through hierarchical clustering. Additionally, the table indicates the size of each cluster, reflecting the number of securities grouped within each.

Table 2. Cluster centroids: annualized mean and standard deviation

Cluster	Cluster Size	Annualized Mean Return	Annualized Standard Deviation
1	4	0.55	0.84
2	1	0.09	0.59
3	20	0.16	0.29
4	21	0.32	0.39
5	34	0.10	0.40
6	50	0.10	0.42

Note: * estimations based on the Yahoo Finance data (Yahoo, 2024).

The pattern draws from the cluster sizes, and their respective metrics can be described as follows:

Cluster 1 comprises only 4 assets. It exhibits the highest annualized mean return at 55.01% and the highest volatility at 84.03%. These assets represent high-risk, high-return investments.

Cluster 2 is the smallest, with just one asset, with a moderate return of 9.43% and a still high volatility of 59.15%. This unusual cluster represents only the LUMN security, which may suggest that LUMN's price movements and risk-return characteristics do not align closely with the rest of the dataset.

Indeed, such statistically non-standard assets (like clusters 1 and 2) cannot be identified due to classical analysis since classical portfolio optimization excludes significant outliers in behavior, including due to the use of a linearized correlation matrix. Therefore, such assets as those presented in Cluster 1 and especially 2 require separate analysis.

Clusters 3 and 4 represent mid-sized groupings of 20 and 21 assets. These clusters achieve average returns of 15.68% and 31.75%, respectively, with 29.15% and 39.49% volatilities, suggesting they contain assets with balanced risk-return profiles.

Clusters 5 and 6, the most prominent groups with 34 and 50 assets, respectively, show relatively low mean returns of 10.32% and 10.49%, coupled with volatilities of 40.35% and 41.53%. These clusters likely represent lower-risk, more stable investments.

http://dx.doi.org/10.21511/imfi.22(1).2025.12

Overall, the clustering results show that smaller clusters tend to group assets with higher volatility in performance, while larger clusters are composed of more consistent, stable assets. These distinctions provide valuable insights for portfolio managers aiming to balance risk and return through diversification.

The comparative analysis of the data presented in Tables 1 and 2 further reveals that the centroids of the clusters obtained through clustering analysis, incorporating features beyond those tied exclusively to classical optimization tasks, tend to exhibit lower average return levels compared to the return of the portfolio optimized through Sharpe ratio optimization. Consequently, the H1, which applied machine learning models consistently resulted in more effective portfolio returns than classical statistical methods, was not proved.

This result is fundamentally linked to the inherent differences in the methodologies. Optimization models based on Markowitz principles are explicitly designed to achieve the best possible optimality criterion, such as maximizing returns while minimizing risk. Conversely, clustering analysis in this context aims to identify broader distinctions among the assets rather than focusing solely on achieving an optimal return or diversification ratio.

In this case, clustering analysis highlights a broader spectrum of heterogeneity in asset characteristics instead of focusing exclusively on the diversification ratio calculated based on volatilities. This broader analytical perspective underscores the complementary role of clustering in understanding asset diversity and its implications for portfolio construction, albeit with trade-offs regarding direct return maximization.

Let us proceed and conduct a comparative analysis of the results of classical portfolio optimization and class analysis. Table 3 provides a comparative study of the portfolio structure derived from Sharpe ratio optimization and clustering-based groupings of assets. Each cluster highlights the assets that belong to the optimized portfolio (with weights exceeding 0) and the top three financial assets from the cluster ranked by Sharpe ratio.

The comparison of the optimized portfolio structure and the clustering-based analysis reveals several key findings. For Cluster 1, the assets in the portfolio (TSLA, MARA, MSTR) align closely with the top performers ranked by the Sharpe ratio. This consistency suggests that the clustering method captures assets with strong individual rates of return within this group. In contrast, as noted earlier, Cluster 2 includes LUMN and has no assets overlapping with the optimized portfolio. This discrepancy highlights the unique characteristics of LUMN that are prioritized by clustering but not by portfolio optimization. Cluster 3 shows no overlap between the optimized portfolio and the clustering results, with prominent assets like MSFT, TSM, and V identified by Sharpe ratio ranking within the cluster but excluded from the optimized portfolio. This again confirms the uniqueness of the approach based on machine learning methods. Clusters 4 and 5 exhibit significant alignment, with key assets like NVDA, AVGO, CELH, WMT, and MO appearing in both the optimized portfolio and the clustering analysis. This suggests these clusters contain assets that are recognized well by both methods. In Cluster 6, the clustering analysis identifies assets like UMC, KO, and STLA as top performers by Sharpe ratio, while the portfolio optimization includes KGC, BTG, IAG, and NEM. This divergence reflects differing priorities between the two methods and requires a

Table 3. Cluster-based analysis and portfolio comparison of selected assets

Cluster Number	Assets in Portfolio	Top 3 Assets by Sharpe Rate
1	TSLA, MARA, MSTR	TSLA, MSTR, MARA
2	–	LUMN
3	–	MSFT, TSM, V
4	NVDA, AVGO, PANW, IBN, DXCM, SMCI, CELH, KDP	NVDA, AVGO, CELH
5	WMT, MO, K	WMT, MO, SCHW
6	KGC, BTG, IAG, NEM	UMC, KO, STLA

Note: * estimations based on the Yahoo Finance data (Yahoo, 2024).

detailed study. Overall, the results demonstrate both alignment and divergence between the two approaches, providing a richer understanding of the asset characteristics and their potential contributions to portfolio performance.

The results of Table 3 underscore the need to rethink and expand the classical concept of diversification. Hypothesis H2, "Diversification of a financial investment portfolio can be measured not only by volatility indicators as a variation of returns," is proved by redefining diversification and requires introducing clustering-based metrics.

The comparison of cluster analysis and portfolio optimization results gives additional insights beyond the quantitative outcomes obtained. While fundamental in evaluating financial investments, the classical concept of financial asset diversification (Markowitz, 1952) requires further refinement. The significant relationship between return and risk is decisive in investment management. However, the idea of asset diversification must be expanded to include distinctions not only in terms of risks and mutual dynamics but also across a broader range of essential factors.

For instance, the behavior of market participants may be influenced not solely by the relationship between return and risk but also by the potential for asset price growth or decline, the presence of unique factors, and metrics derived from technical and graphical analysis. These considerations suggest that financial asset and portfolio diversification encompasses a much broader scope of characteristics than traditionally addressed.

For instance, the traditional diversification ratio can consider cluster diversity. The cluster-related diversification ratio provides a more structured measure of how variance is distributed within and across clusters by all features. Cluster diversity evaluates how well-separated and compact clusters remain internally, using total between-cluster and total within-cluster variance. This metric can be calculated as follows:

$$Diversification\ ratio = \frac{Between\ cluster\ variance}{Between\ cluster\ variance + Within\ cluster\ variance}. \quad (1)$$

Additionally, a weighted approach can be incorporated to consider the cluster or asset weights, providing a more nuanced assessment of diversification.

Thus, diversification is the strategic allocation of financial assets within a portfolio to reduce overall risk while maximizing potential returns. It incorporates price growth or decline trends, technical, graphical, and fundamental analysis indicators, and other unique asset-specific influences.

4. DISCUSSION

The results highlight the potential of clustering analysis tools in enhancing portfolio diversification. Compared to traditional financial models, the clustering approach offered a broader perspective by grouping assets based on multidimensional characteristics rather than solely relying on volatility and return. These findings align with previous studies, such as those by López de Prado (2016) and Owen (2023), who emphasized the advantages of clustering in identifying hidden patterns and improving diversification. However, there are notable differences in the application of clustering techniques, particularly in the selection of features and the interpretability of results.

One significant challenge in the paper was determining the optimal number of clusters. Previous studies, including those by Jain and Jain (2019), have not shown a robust technique to define the cluster numbers. In the paper, this definition depends on Elbow or Silhouette Score analysis methods. Furthermore, the observed cluster structures were sensitive to the choice of features, reflecting the complexity of financial asset behavior. The reliance on classical risk-return metrics and technical indicators influenced the clustering outcomes, focusing on capturing assets' statistical and behavioral profiles. However, it may have overlooked other dimensions, such as fundamental factors or broader macroeconomic influences.

The findings suggest that expanding the feature set for clustering, incorporating elements such as

http://dx.doi.org/10.21511/imfi.22(1).2025.12

graphical analysis, market sentiment, and fundamental metrics, could provide a more comprehensive understanding of asset relationships. Previous research (e.g., Jain & Jain, 2019; López de Prado, 2016; Owen, 2023) demonstrated the value of integrating many different metrics, mainly from a traditional point of view. Additionally, the clustering results underscored the importance of stability and validation. The observed variability in clustering outcomes with slight feature changes highlights the need for robust cross-validation techniques.

CONCLUSION

The paper underscores the growing importance of machine learning (ML) in modern investment management. With financial markets becoming increasingly complex and the volume of available information expanding, relying solely on traditional approaches is no longer sufficient. ML models have proven essential for making timely, informed decisions and uncovering new patterns in financial data, and many publications in this area confirm this.

This study aimed to explore the potential of clustering analysis in redefining portfolio diversification by going beyond traditional metrics such as volatility. The findings demonstrate that clustering analysis offers valuable results by identifying meaningful groupings of assets based on multidimensional characteristics, including graphical, technical, fundamental analysis indicators, and statistical metrics. These insights provide a more reliable understanding of asset behavior and contribute to the broader concept of diversification.

While the classical portfolio optimization method based on Markowitz principles remains a gold standard, it has limitations. It is highly effective in mathematical optimization but does not fully account for the nuances of market behavior or the rapidly growing pool of management data. This creates opportunities for integrating additional methods to address these gaps. The cluster analysis identified six meaningful clusters within the global market's top 130 securities. This method went beyond classical metrics like return and risk by incorporating technical analysis indicators and autocorrelation coefficients, capturing a more dynamic picture of asset behavior. Due to its inherent nature, cluster analysis results don't outperform strict classical optimization models regarding portfolio return. However, further development and application across diverse datasets could significantly improve its effectiveness. In conclusion, this study demonstrates the significant potential of clustering analysis in optimizing portfolio structures. By leveraging clustering-based approaches, investors can identify distinct asset groupings that enhance diversification metrics. Furthermore, this optimization framework bridges the gap between classical financial models and modern, data-driven strategies, offering a dynamic approach to portfolio construction. Further research will address challenges such as cluster interpretability, weight allocation, and the integration of clustering outputs into traditional optimization models.

DIRECTIONS FOR FURTHER DEVELOPMENT

The paper underscores the need to rethink and expand the classical concept of diversification. The cluster-related diversification ratios may provide a more structured measure of how variance is distributed within and across clusters by all features. These metrics are derived from clustering analysis techniques, providing a more comprehensive diversification perspective. Consequently, the diversification of an investment portfolio should not only balance risk and return but also consider broader factors such as behavioral patterns, technical and fundamental indicators, and descriptive statistical variables. By incorporating these dimensions, investors can better understand their portfolios and make more informed decisions.

An important direction for further development is the application of dynamic clustering techniques. Unlike static analyses, which consider only a snapshot of variables, dynamic clustering examines patterns of change over time, such as price trends or return movements. This approach could enhance the ability to capture the evolving nature of financial markets and provide more actionable insights for portfolio management. Another critical technical aspect that warrants further investigation is cross-validation methods. This step is crucial for validating machine learning models and cluster analysis and ensuring the applicability of classical financial models in rapidly changing markets. In the modern world of big data, where markets shift at an unprecedented pace, identifying stable quantitative patterns is vital. Without this validation, any model risks quickly becoming outdated and ineffective.

AUTHOR CONTRIBUTIONS

Author contributions Mykhailo Kuzheliev, Dmytro Zherlitsyn.
Conceptualization: Dmytro Zherlitsyn.
Data curation: Mykhailo Kuzheliev.
Formal analysis: Ihor Rekunenko, Alina Nechyporenko, Sergii Stabias.
Investigation: Mykhailo Kuzheliev, Dmytro Zherlitsyn, Alina Nechyporenko.
Methodology: Mykhailo Kuzheliev, Dmytro Zherlitsyn, Ihor Rekunenko.
Project administration: Mykhailo Kuzheliev, Ihor Rekunenko.
Resources: Alina Nechyporenko, Sergii Stabias.
Software: Dmytro Zherlitsyn, Ihor Rekunenko, Sergii Stabias.
Supervision: Mykhailo Kuzheliev.
Validation: Ihor Rekunenko, Sergii Stabias.
Visualization: Dmytro Zherlitsyn, Alina Nechyporenko.
Writing – original draft: Mykhailo Kuzheliev, Dmytro Zherlitsyn, Ihor Rekunenko.
Writing – review & editing: Alina Nechyporenko, Sergii Stabias.

REFERENCES

1. Agudelo Aguirre, A. A., Rojas Medina, R. A., & Duque Méndez, N. D. (2020). Machine learning applied in the stock market through the Moving Average Convergence Divergence (MACD) indicator. *Investment Management and Financial Innovations, 17*(4), 44-60. https://doi.org/10.21511/imfi.17(4).2020.05
2. Aiche, A., Winer, Z., & Cohen, G. (2024). Constructing Cybersecurity Stocks Portfolio Using AI. *Forecasting, 6*(4), 1065-1077. https://doi.org/10.3390/forecast6040053
3. Apalkova, V., Tsyganov, S., Meshko, N., Tsyganova, N., & Apalkov, S. (2022). Evaluation models for the impact of pricing factor on environmental performance in different countries. *Problems and Perspectives in Management, 20*(2), 135-148. https://doi.org/10.21511/ppm.20(2).2022.12
4. Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2021). Machine learning in finance: A topic modeling approach. *European Financial Management*. https://doi.org/10.1111/eufm.12326
5. Babenko, V., Panchyshyn, A., Zomchak, L., Nehrey, M., Artym-Drohomyretska, Z., & Lahotskyi, T. (2021). Classical machine learning methods in economics research: Macro and micro level examples. *WSEAS Transactions on Business and Economics, 18,* 209-217. https://doi.org/10.37394/23207.2021.18.22
6. Bhama, V. (2024). Does an increase in portfolio volatility create more returns? Evidence from India. *Investment Management and Financial Innovations, 21*(2), 345-354. https://doi.org/10.21511/imfi.21(2).2024.28
7. Clarissa, A., & Koesrindartoto, D. P. (2024). Strategic portfolio rebalancing: Integrating predictive models and adaptive optimization objectives in a dynamic market. *Investment Management and Financial Innovations, 21*(3), 304-316. https://doi.org/10.21511/imfi.21(3).2024.25
8. Derbentsev, V., Datsenko, N., Babenko, V., Pushko, O., & Pursky, O. (2021). Forecasting cryptocurrency prices using ensembles-based machine learning approach. In *2020 IEEE International Conference on Problems of Infocommunications Science and Technology (PIC S&T) Proceedings* (pp. 707-712). https://doi.org/10.1109/PICST51311.2020.9468090
9. Fantazzini, D., & Zimin, S. (2020). A multivariate approach for the simultaneous modelling of market risk and credit risk for cryptocur-

http://dx.doi.org/10.21511/imfi.22(1).2025.12

rencies. *Journal of Industrial and Business Economics, 47*(1), 19-69. https://doi.org/10.1007/s40812-019-00136-8

10. Feng, X., von Mettenheim, H.-J., Sermpinis, G., & Stasinakis, C. (2024). Sustainable portfolio construction via machine learning: ESG, SDG, and sentiment. *European Financial Management.* https://doi.org/10.1111/eufm.12531
11. Gallastegui, L. M. G., Forradellas, R. R., & Alonso, S. L. N. (2024). Applying advanced sentiment analysis for strategic marketing insights: A case study of BBVA using machine learning techniques. *Innovative Marketing, 20*(2), 100-115. https://doi.org/10.21511/im.20(2).2024.09
12. Glazunova, O., Saiapina, T., Korolchuk, V., Kasatkina, O., & Voloshyna, T. (2021, May 12-14). Digital intelligence of a modern economist: An exploratory case study. Paper presented at the *2nd International Conference on History, Theory and Methodology of Learning (ICHTML).* Kryvyi Rih, Ukraine.
13. Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3-12. https://doi.org/10.1002/asmb.2209
14. Inani, S. K., Pradhan, H., Kumar, S., & Biswas, B. (2024). Navigating the technical analysis in stock markets: Insights from bibliometric and topic modeling approaches. *Investment Management and Financial Innovations, 21*(1), 275-288. https://doi.org/10.21511/imfi.21(1).2024.21
15. Jain, P., & Jain, S. (2019). Can Machine Learning-Based Portfolios Outperform Traditional Risk-Based Portfolios? The Need to Account for Covariance Misspecification. *Risks, 7*(3), 74. https://doi.org/10.3390/risks7030074
16. Korstanje, J. (2021). *Advanced forecasting with Python: With state-of-the-art models including LSTMs, Facebook's Prophet, and Amazon's DeepAR.* Apress. https://doi.org/10.1007/978-1-4842-7150-6
17. Kuzheliev, M., Rekunenko, I., Boldova, A., Zhytar, M., & Stabias, S. (2019). Modeling of structural and temporal characteristics in the corporate securities market of Ukraine. *Investment Management and Financial Innovations, 16*(2), 260-269. http://dx.doi.org/10.21511/imfi.16(2).2019.22
18. Kuzheliev, M., Zherlitsyn, D., Rekunenko, I., Nechyporenko, A., & Nemsadze, G. (2020). The impact of inflation targeting on macroeconomic indicators in Ukraine. *Banks and Bank Systems, 15*(2), 94-104. https://doi.org/10.21511/bbs.15(2).2020.09
19. Leung, M.-F., Jawaid, A., Ip, S.-W., Kwok, C.-H., & Yan, S. (2023). A portfolio recommendation system based on machine learning and big data analytics. *Data Science in Finance and Economics, 3*(2), 152-165. https://doi.org/10.3934/DSFE.2023009
20. Liew, J. K. S., & Mayster, B. (2018). Forecasting ETFs with machine learning algorithms. *Journal of Alternative Investments, 20*(3), 58-78. https://doi.org/10.3905/jai.2018.20.3.058
21. López de Prado, M. (2016). Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management, 42*(4), 59-69. https://doi.org/10.3905/jpm.2016.42.4.059
22. Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77-91. https://doi.org/10.2307/2975974
23. Mints, A. (2017). Classification of tasks of data mining and data processing in the economy. *Baltic Journal of Economic Studies, 3*(3), 47-52. https://doi.org/10.30525/2256-0742/2017-3-3-47-52
24. Owen, S. R. (2023). An analysis of conditional mean-variance portfolio performance using hierarchical clustering. *The Journal of Finance and Data Science, 9,* 100112. https://doi.org/10.1016/j.jfds.2023.100112
25. Pinelis, M., & Ruppert, D. (2022). Machine learning portfolio allocation. *The Journal of Finance and Data Science, 8,* 35-54. https://doi.org/10.1016/j.jfds.2021.12.001
26. Sang, N.M. (2024). Bibliometric insights into the evolution of digital marketing trends. *Innovative Marketing, 20*(2), 1-14. https://doi.org/10.21511/im.20(2).2024.01
27. Viebig, J. (2020). Exuberance in financial markets: Evidence from machine learning algorithms. *Journal of Behavioral Finance, 21*(2), 128-135. https://doi.org/10.1080/15427560.2019.1663849
28. Yahoo! (2024). *Yahoo! Finance Data.* Retrieved from https://finance.yahoo.com
29. Zherlitsyn, D. (2024) *Python for Finance: Data analysis, financial modeling, and portfolio management* (English Edition) (1st ed.). BPB Publications. Retrieved from https://read.kortext.com/reader/pdf/3268854
30. Zmuk, B., & Josic, H. (2020). Forecasting stock market indices using machine learning algorithms. *Interdisciplinary Description of Complex Systems, 18*(4), 471-489. https://doi.org/10.7906/indecs.18.4.7