Stock Clustering Framework using Financial Ratios: A Case Study in the Stock Exchange of Thailand

Kietikul Jearanaitanakij*, Natdanai Poonpon, Chanidapa Wongtep, Kittaporn Buriyameathakul and Artitaya Pimsupaporn

Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Lat Krabang, Lat Krabang, Bangkok, 10520, Thailand

***Corresponding Author E-mail:** kietikul.je@kmitl.ac.th

Received: Jul 30, 2025; **Revised:** Oct 09, 2025; **Accepted:** Oct 10, 2025

Abstract

Value investors typically seek undervalued stocks that align with specific financial criteria to maximize their margin of safety. However, manually analyzing the financial data of all listed stocks is a time-intensive process. Furthermore, the market price of a target stock may exceed its intrinsic value, introducing potential investment risks. To address these challenges, this study proposes a stock clustering framework that groups equities based on financial ratio similarity. The proposed framework is designed to streamline the investment decision-making process by recommending stocks with comparable financial profiles as alternatives to those currently attracting investor interest but that may already be overvalued. Multiple clustering algorithms are evaluated to determine the most effective grouping strategy. Empirical back testing using four years of data from the Stock Exchange of Thailand reveals that the Gaussian Mixture Model (GMM) achieves the highest composite performance metric among the tested methods. Additionally, the HDBSCAN algorithm is employed to detect and exclude outlier stocks, thereby enhancing the reliability of the clustering results.

Keywords: Stock clustering, Composite index, Financial ratios, Stock exchange of Thailand, Investment

1. Introduction

Investing in stocks is a strategic financial move due to their potential for high returns. It offers the benefit of dividend income, which can supplement other earnings, and allows for portfolio diversification as different stocks react uniquely to market conditions. When you buy shares, you gain a small ownership stake in the company, sharing in its success. Stocks are generally liquid, allowing for flexibility in buying or selling on any business day. They also offer protection against inflation, as investments in equities can increase in value over time due to rising company revenues and profits. However, investing in stocks also carries risks, so thorough research and possibly advice from a financial advisor are necessary. Different techniques make analyzing stocks more convenient. Their literature reviews are described as follows.

Wang et al. [1] introduced the Hurst exponent to scrutinize the long-memory characteristics of the China stock market. The outcomes of the experiments shed light on the market's dynamics. A Hurst exponent nearing 0.5 implies that stock prices follow a random walk pattern. Conversely, a Hurst exponent substantially above 0.5 indicates a long memory, suggesting that historical changes in stock prices have a significant impact on future price fluctuations. Jearanaitanakij and Passaya [2] introduced a framework that utilizes a convolutional neural network and candlestick patterns to forecast short-term stock trends. The methodology was tested using candlestick pattern images gathered from various stocks listed on the Stock Exchange of Thailand (SET). Each image encapsulates between six and twelve consecutive candlesticks. The findings suggest that the proposed technique can accurately anticipate the short-term trend for most stocks with acceptable accuracy.

Liu et al. [3] forecasted the trends of 50 stocks on the Shanghai Stock Exchange. They integrated an LSTM model with a stock attention mechanism and incorporated supplementary data from the stock cluster to predict the trajectory of stock prices. Their findings revealed that the inclusion of stock cluster data notably enhanced the performance of the LSTM model on the dataset obtained from the Wind Financial Terminal. Lee et al. [4] implemented a Deep Q-Network, complemented by a Convolutional Neural Network function approximator that uses stock chart images as input, to predict global stock market trends. The model proved profitable not only in the country's stock market where it was initially trained, but also demonstrated profitability in various global stock markets. This indicates the model's potential applicability to a wide range of global markets, extending beyond its original training ground. Indriyanti and Dhini [5] employed a model, grounded in the Gaussian Mixture Model and the Expectation-Maximization (EM) algorithm, to cluster high-dimensional stock data. The study juxtaposes the efficacy of their method and its amalgamation with PCA in clustering high-dimensional time series data. The findings reveal that although PCA enhances the time efficiency of model construction, their method outperforms the combined approach with PCA in handling high-dimensional data.

Moedjahedy et al. [6] predict the stock prices of five firms in the telecommunications industry using the Gaussian Process and SMOreg algorithms. Their findings indicated that the SMOreg algorithm

outperformed the Gaussian Process in terms of RMSE, MAPE, and MBE on a dataset gathered three years prior. Patil and Joshi [7] devised a method that integrates clustering and candlestick patterns for forecasting stock market movements. Their model enhances clustering with a widely used technical analysis technique known as the candlestick method. They employ rough set-based clustering to categorize similar stocks or trends, while the candlestick method is utilized for analyzing and predicting upcoming trends. Their model demonstrates superior performance compared to models that solely use clustering or candlestick techniques.

Shirota and Murakami [8] utilized long-term time series data clustering for forecasting stock prices and selecting portfolios. Their study employed two clustering techniques, k-Shape and k-means with Dynamic Time Warping (DTW) distance measure, to examine stock data from the leading 129 global electronics manufacturers spanning from 2018 to 2020. The results from the k-Shape clustering highlighted distinct impacts on various countries' stock markets due to the COVID-19 pandemic. Interestingly, each of the eight clusters is composed of companies from a single country, implying that investors or their algorithms might favor investing in companies based on their country of origin rather than the performance of the individual company. Wang et al. [9] combined morphological similarity distance (MSD) and k-means clustering to identify stocks with similar characteristics. Subsequently, an online learning model known as Hierarchical Temporal Memory (HTM) was employed to discern patterns from these similar stocks and forecast their prices. The results demonstrated that this method, which learns from similar stock patterns, exhibits superior prediction accuracy compared to the HTM model that does not incorporate such learning. Naik and Mohan [10] proposed a model based on the Hybrid Feature Selection (HFS) technique, which initially eliminates non-essential financial parameters from the stock data. Subsequently, the Naive Bayes method is employed to categorize stocks with strong fundamentals, and the Relative Strength Index (RSI) is used to detect bubbles in stock prices. Crisis points in stock prices are identified using moving average statistics. Performance evaluation on an Indian stock dataset revealed that the HFS-based XGBoost method outperforms the HFS-based DNN method in predicting stock crises. Leangarun et al. [11] proposed an innovative method for identifying stock price manipulation. Their technique employs deep unsupervised learning to instruct neural networks to discern standard trading behaviors depicted in a limit order book. Trading activities that deviate from these learned patterns are flagged as manipulative. Experimental outcomes, based on six legally prosecuted manipulation instances in SET, reveal that both autoencoder and generative adversarial networks can accurately detect five out of the six cases, maintaining a low rate of false positives. Ploysuwan and Pravithana [12] introduced a method for grouping movements in stock prices using self-supervised learning and continuous wavelet transforms. They tackled the complexities of contemporary clustering algorithms dealing with stock similarity. Their technique employs self-supervised learning and continuous wavelet transform to identify similarities among stocks. This strategy improves the detection of stock similarities, offering crucial perspectives for financial analysis and investment planning.

Kim et al. [13] presented an innovative framework, known as ASA, for the autonomous selection and allocation of stocks. This framework integrates ranking models with classification and regression models. Regarding stock selection, the framework employs a hybrid approach, utilizing simple graph and hypergraph-based ranking models for relational modeling to pinpoint the most lucrative stocks. A combination of classification and regression models is used to establish the investment ratio for the allocation of stocks. Experimental outcomes on the Standard & Poor's 500 index reveal that ASA attains a compounded annual growth rate significantly superior to the second-best performing method. Wang [14] integrated Bidirectional Long Short-Term Memory (BiLSTM) networks with an enhanced Transformer structure and a Temporal Convolution Network (TCN) to boost the precision and stability of stock price forecasts. The Transformer model was refined and adapted for stock price prediction by incorporating TCN, which can grasp sequence dependencies and enhance the model's ability to generalize. The BiLSTM model was employed to capture bidirectional information in sequences. The experimental outcomes suggest that this method excels in predicting stock prices, demonstrating high accuracy and a strong ability to generalize. Li et al. [15] presented a novel hybrid model for stock price prediction. The model employs complete ensemble empirical mode decomposition with adaptive noise for the initial breakdown of the stock price time series. Sub-series that share similar sample entropy from the decomposition are grouped using the K-means clustering technique. Each sub-series is then individually forecasted using the gated recurrent unit (GRU) model. The final predictions are derived by combining the outcomes of these individual forecasts. The model has been tested on three distinct stock markets and has demonstrated superior performance compared to other forecasting methods across all stock indices. Chakravorty and Elsayed [16] investigate the effectiveness of various machine learning algorithms—specifically, decision trees, random forests, support vector machines (SVMs), and K-means clustering—for predicting stock prices using insider trading data, with a focus on Tesla stock from 2020 to 2023. They apply Recursive Feature Elimination (RFE) and feature importance analysis to optimize model performance, finding that SVM with a radial basis function (RBF) kernel delivers the highest accuracy, albeit with greater computational cost. The research highlights the potential of integrating insider trading signals into predictive models to enhance financial forecasting.

It would be beneficial if we could identify stocks with similar financial ratios to a stock we are interested in, whose price may still be below its intrinsic value. In this research, we propose a stock clustering framework based on fourteen financial ratios. We employ four clustering algorithms (Affinity Propagation, Agglomerative, Gaussian Mixture, and K-Means) to find stock clusters and exclude stocks as suggested as noise or outliers by the HDBSCAN algorithm. Investors can save time by considering only stocks in the same group as their interested stock. The experimental results on public data for the past four years from the stock exchange of Thailand reveal that the Gaussian Mixture algorithm possesses the best composite index. After analyzing selected stock examples in detail, we found that the proposed framework can group stocks with similar financial ratios in the same cluster.

The remainder of this paper is structured as follows: Section 1 imparts the essential knowledge required to comprehend this study. Section 2 outlines the stock dataset and the framework proposed. Section 3 describes the proposed stock clustering framework and a novel composite index. Section 4 encompasses the experiments and a comparison of all clustering techniques, accompanied by discussions. Lastly, the conclusion and the suggested clustering algorithm, along with sample clusters of stocks frequently categorized together, are described in Section 5.

2. Fundamental Knowledge

To provide readers with the essential background information and context, the foundational knowledge necessary for understanding our research is briefly explained in this section.

2.1 Financial Ratios

The fundamentals of financial ratios [17] used in our stock clustering experiments are listed in **Table 1**. These ratios play a crucial role and are commonly employed by value investors to assess a company's financial stability, profitability, operational efficiency, and overall valuation. They offer a multidimensional analysis that covers valuation, profitability, financial stability, operational efficiency, and cash flow health. This makes them not only important but also sufficient for making informed investment decisions.

Table 1 Financial ratios and their descriptions

Financial ratio	Description
Price to Earning (P/E)	The P/E ratio helps investors determine the market value of a stock compared to the company's earnings.
Price to Book Value (P/Bv)	The P/Bv ratio indicates whether a stock is over or undervalued by comparing the market's valuation to the company's actual worth.
Book Value Per Share	The value of a company's equity on a per-share basis.
Return on Asset (ROA)	ROA measures how efficiently a company uses its assets to generate profit.
Return on Equity (ROE)	ROE measures a company's profitability by revealing how much profit a company generates with the money shareholders have invested.
Net Profit Margin	The net profit margin shows the ratio of revenue that remains profit after all expenses are deducted. It indicates how well a company controls its costs.
Dividend Yield	Dividend yield shows how much a company pays out in dividends each year relative to its stock price.
Debt-to-Equity (D/E)	The D/E ratio indicates the proportion of debt and equity used to finance the company's assets.
Total Assets Turnover	Total assets turnover measures how efficiently a company uses its assets to generate sales. It indicates the effectiveness of asset utilization.
Operating Cash Flow Margin	The ratio of cash generated from a company's operating activities relative to its total revenue.
Investing CF Margin	The ratio of cash flow generated or used by a company's investing activities relative to its total revenue.
Financing CF Margin	The ratio of cash flow generated or used by a company's financing activities relative to its total revenue.
Earnings Before Interest and Taxes Margin (EBIT Margin)	EBIT Margin indicates how much revenue is left over after operating expenses, excluding interest and taxes, have been deducted. This margin helps investors and analysts understand how efficiently a company is being managed and how well it is generating profits from its operations.
Fixed Asset Turnover	This ratio measures a company's efficiency in utilizing its fixed assets to generate revenue.

2.2 K-Means Clustering

K-means clustering is a fundamental unsupervised learning algorithm used in data mining and machine learning to partition a dataset into (K) distinct clusters. The algorithm aims to minimize the variance within each cluster by iteratively assigning data points to the nearest cluster centroid and then recalculating the centroids based on the mean of the designated points.

This process continues until the centroids stabilize, ensuring that the data points within each cluster are as similar as possible while being distinct from those in other clusters. The effectiveness of K-Means Clustering lies in its simplicity and computational efficiency, making it suitable for large datasets. However, it is sensitive to the initial placement of centroids and can be affected by outliers. The algorithm is widely applied in various fields, including market segmentation, image compression, and document clustering, due to its ability to uncover hidden patterns in data [18].

2.3 Affinity Propagation Clustering

Affinity Propagation Clustering is an innovative clustering algorithm introduced by Frey and Dueck [19]. Unlike traditional clustering methods that require the number of clusters to be specified in advance, Affinity Propagation identifies clusters by simultaneously considering all data points as potential exemplars and exchanging real-valued messages between data points. The algorithm operates by iteratively updating two types of messages: "responsibility," which reflects how well-suited a data point is to serve as an exemplar for another point, and "availability," which indicates the appropriateness for a point to choose another as its exemplar. This message-passing process continues until convergence, resulting in the selection of exemplars that best represent the data clusters. Affinity Propagation is particularly effective in discovering clusters of varying sizes and shapes. It is robust to noise, making it suitable for various applications, including image processing and bioinformatics.

2.4 Agglomerative Clustering

Agglomerative Clustering is a hierarchical clustering method that builds nested clusters by successively merging or splitting them. This algorithm starts with each data point as an individual cluster and iteratively merges the closest pairs of clusters until all points are contained in a single cluster or a predefined number of clusters is reached. The proximity between clusters can be measured using various linkage criteria, such as single linkage (minimum distance), complete linkage (maximum distance), or average linkage (mean distance). Agglomerative Clustering is particularly useful for revealing the hierarchical structure of data, making it applicable in fields such as bioinformatics, image analysis, and social network analysis. One of the earliest formalizations of this method was presented by Sokal and Michener [20], who applied it to biological taxonomy to classify organisms based on their characteristics.

2.5 Gaussian Mixture Clustering

The Gaussian Mixture (GM) [21] is a probabilistic model that assumes all the data points are generated from a mixture of several Gaussian distributions with unknown parameters. This model is widely used in various fields such as machine learning, pattern recognition, and statistical data analysis due to its flexibility and ability to model complex data distributions. The Expectation-Maximization (EM) algorithm is typically employed to estimate the parameters of the Gaussian mixtures, iteratively improving the likelihood of the observed data under the model. The GMM is particularly useful in clustering applications, where it can identify subpopulations within an overall population without requiring labeled data.

2.6 Hierarchical Density-Based Spatial Clustering of Applications with Noise

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [22] is an advanced clustering algorithm that extends DBSCAN by converting it into a hierarchical clustering algorithm. HDBSCAN builds a hierarchy of clusters and then condenses this hierarchy based on the stability of clusters, allowing for the extraction of clusters at varying densities. The algorithm operates by first constructing a minimum spanning tree of the data points, which is then used to create a hierarchy of clusters. This hierarchy is pruned to form the final clustering, where clusters are defined by their stability over various scales. HDBSCAN is particularly effective in identifying clusters of varying shapes and densities and is robust to noise, making it suitable for complex datasets. Its ability to handle varying densities and hierarchical nature provides a more nuanced clustering than traditional methods.

2.7 Silhouette Score

The Silhouette Score [23] is a metric used to evaluate the quality of clusters in a clustering analysis. It measures how similar an object is to its cluster compared to others, thus providing a combined assessment of cohesion and separation. The score ranges from -1 to 1, where a higher value indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters. A score close to 1 suggests that the clusters are well-defined, while a score near 0 indicates overlapping clusters, and negative values imply that the objects might have been assigned to the wrong clusters. The Silhouette Score is calculated by averaging the silhouette coefficient of all data points, where the silhouette coefficient for a single point is defined as the difference between the mean intra-cluster distance and the mean nearest-cluster distance, divided by the maximum of these two distances. The major limitations of the Silhouette Score include the sensitivity to noise and outliers, and the poor performance with clusters of varying shapes and densities.

2.8 Calinski-Harabasz Index

The Calinski-Harabasz Index (CHI) [24], also called the Variance Ratio Criterion, serves as an internal evaluation metric for assessing the quality of clustering outcomes. It quantifies the ratio of inter-cluster separation to intra-cluster dispersion, adjusting for their respective degrees of freedom. Elevated CHI values signify superior cluster definition, reflecting well-separated and compact clusters. This index is especially advantageous for identifying the optimal

number of clusters in clustering algorithms such as k-means. Besides the sensitivity to noise, similar to the limitation of the Silhouette Score, CHI also suffers in accurately overestimating the optimal number of clusters since it favors solutions with more clusters.

2.9 Davies-Bouldin Index

The Davies-Bouldin Index (DBI) [25] is an internal evaluation metric for clustering quality. The DBI quantifies the ratio of intra-cluster dispersion to inter-cluster separation, with lower values indicating superior clustering quality. This metric is particularly advantageous for identifying the optimal number of clusters in clustering algorithms such as k-means, as it accounts for individual clusters' compactness and distinctness. However, the DBI assumes that clusters are spherical and equally sized. This assumption can be problematic when dealing with clusters of varying shapes and sizes, as it may not accurately reflect the true clustering quality.

2.10 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) [26] is a powerful technique for visualizing high-dimensional data by mapping each data point to a location in a two or three-dimensional space. t-SNE works by constructing a probability distribution over pairs of high-dimensional objects, assigning higher probabilities to similar objects and lower probabilities to dissimilar ones. It then defines a similar probability distribution in the lower-dimensional space and minimizes the Kullback-Leibler divergence between these two distributions to preserve the relative distances between points. This method is particularly useful for visualizing complex datasets in fields such as genomics, natural language processing, and bioinformatics.

3. Stock Clustering Framework

To measure the performance of the clustering algorithms, we formulate a new composite index by averaging three normalized values of well-known clustering indices (Silhouette Score, Calinski-Harabasz, and Davies-Bouldin) as shown in Equation (1). $\widehat{Sil}$, $\widehat{CHI}$, and $\widehat{DBI}$ are normalized values of Silhouette, Calinski-Harabasz, and Davies-Bouldin, respectively. These 3 indices are normalized into the range between 0 and 1 before being taken into an arithmetic average to yield the range 0 and 1 of the composite index. The higher the composite index value, the better the data are clustered.

$$Composite\ Index = \frac{\widehat{Sil}+\widehat{CHI}+\widehat{DBI}}{3} \quad (1)$$

A composite index offers a comprehensive and robust evaluation of clustering performance by balancing the assessment across multiple dimensions. The Silhouette Score measures cohesion and separation, the Calinski-Harabasz Index evaluates cluster dispersion, and the Davies-Bouldin Index considers intra-cluster and inter-cluster distances. Combining these indices provides a more reliable and nuanced assessment, reducing the risk of skewed results from any single metric and easing decision-making for selecting the optimal clustering algorithm and parameters.

Figure 1 illustrates the proposed stock clustering framework. We employ five prominent clustering algorithms—Affinity Propagation (AP), Agglomerative (AG), Gaussian Mixture (GM), K-Means, and HDBSCAN—to categorize stocks based on their 14 financial ratios. The clustering result with the best composite index value is chosen. Finally, stocks detected as noise and outliers suggested by HDBSCAN are excluded from their clusters.

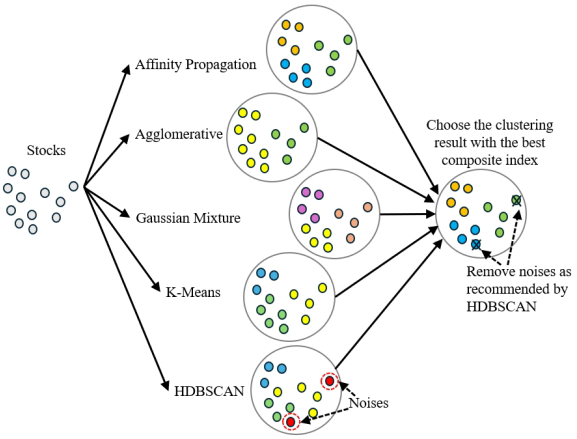

Figure 1 Stock clustering framework

4. Experimental Results and Comparisons

This section presents the dataset utilized in the research experiments, outlines the research objectives, and discusses the experimental results that support these objectives.

4.1 Stock Exchange of Thailand Dataset

We gathered a back-testing dataset of 14 financial ratios in **Table 1** using the SETSmart API. The dataset is collected from SET during 2020 and 2023. To avoid sensitivity to the scale of the data, all financial ratios are rescaled in a range between 0 and 1. In addition, stocks with any missing financial ratios are excluded from the clustering process. We do not replace the missing financial ratios with their means or medians because this can mislead about the financial characteristics of a stock. For example, a missing value of P/E means a company is experiencing operational losses. Using the average as the P/E ratio will incorrectly change the company's financial status from a loss to a profit.

4.2 Research objectives

Our research has two objectives. First, we identify the most suitable clustering algorithm for stock clustering in Thailand based on the fundamental financial ratios. Second, we identify groups of stocks whose financial ratios are similar and frequently clustered in the same group, considering the past data.

For easy understanding, we select only stocks listed in SET50 for 4 consecutive years in the last quarter of 2020-2023. Additionally, these stocks have net

operating profit. 29 selected stocks that met our criteria are alphabetically listed as follows: ADVANC, BDMS, BEM, BH, BTS, CBG, CPALL, CPF, CPN, CRC, EA, EGCO, GLOBAL, GPSC, GULF, HMPRO, INTUCH, IVL, LH, MINT, MTC, OSP, PTT, PTTEP, PTTGC, SCC, SCGP, TOP, and TU. In addition, to facilitate the visualization of the results, we reduced the 14 dimensions of the stock data to 2 dimensions using the t-SNE dimensionality reduction technique.

Although the stock selection in this study focuses exclusively on firms with consistent profitability during the period of interest–an approach that aligns with the preferences of conservative investors who prioritize financial stability–this methodology does not encompass alternative investment strategies. For instance, risk-tolerant investors may intentionally target companies with negative earnings, as such firms can occasionally offer high-return opportunities despite their elevated risk profiles.

4.3 Result for the first objective: The most suitable clustering algorithm

We experiment with a grid search of each clustering algorithm to find the parameter values that produce the highest composite index. **Table 2** lists the best parameters for 5 algorithms and their best composite indices.

Table 2 Best parameters returned from the grid search of each clustering algorithm

Parameter	AP	AG	GM	K-Means	HDBSCAN
init (k-means++, random)	n/a	n/a	n/a	random	n/a
n_init (1-20)	n/a	n/a	1	10	n/a
max_iter (100-1500)	100	n/a	100	1200	n/a
tol (0.0001, 0.001, 0.01)	n/a	n/a	0.001	0.001	n/a
algorithm K-Means: lloyd, elkan, auto HDBSCAN: brute, kd-tree, ball-tree, auto	n/a	n/a	n/a	Lloyd	auto
n_clusters (2-5)	n/a	2	n/a	n/a	n/a
min_cluster_size (2-8)	n/a	n/a	n/a	n/a	5
min_samples (2-8)	n/a	n/a	n/a	n/a	3
cluster_selection_epsilon (0-1)	n/a	n/a	n/a	n/a	1
metric (Euclidean, Manhattan)	n/a	Manhattan	n/a	n/a	Euclidean
affinity (Euclidean, Manhattan)	Euclidean	n/a	n/a	n/a	n/a
damping(0.5-0.9)	0.5	n/a	n/a	n/a	n/a
convergence_iter (10-40)	30	n/a	n/a	n/a	n/a
Preference (-250 to -1000)	-500	n/a	n/a	n/a	n/a
n_components (2-7)	n/a	n/a	6	n/a	n/a
covariance_type (full,tied,diagonal,spherical)	n/a	n/a	spherical	n/a	n/a
reg_covar (1e-8, 1e-6, 1e-3)	n/a	n/a	1e-6	n/a	n/a
Linkage (complete, average)	n/a	complete	n/a	n/a	n/a

Based on the best composite indices in **Table 3**, **Figures 2–5** show the results of 5 clustering algorithms from 2020 to 2023. Stocks of the same color are clustered in the same group. Note that while 14 dimensions (financial ratios) were used during the clustering process, the number of dimensions was reduced to 2 for visualization purposes, to facilitate the interpretation

Table 3 Best composite index for each algorithm

Year	AP	AG	GM	K-Means	HDB SCAN
2020	0.96	0.96	**0.99**	0.92	0.89
2021	**1.00**	0.96	0.94	0.89	0.69
2022	0.94	0.97	**0.99**	0.90	0.90
2023	0.84	0.96	**0.98**	0.79	0.93

.

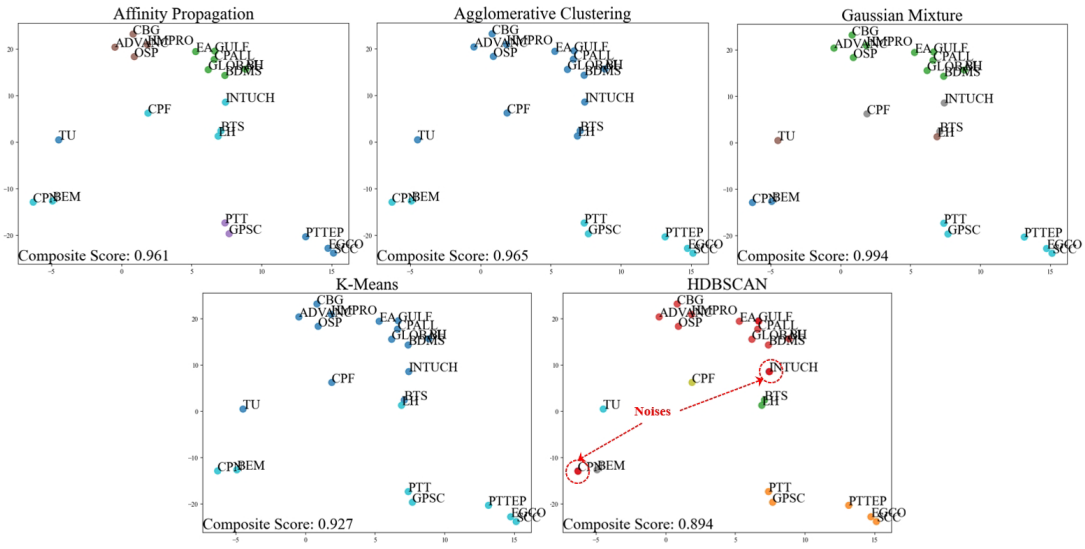

Figure 2 A visualization of clustering results from five algorithms applied to 29 stocks in 2020

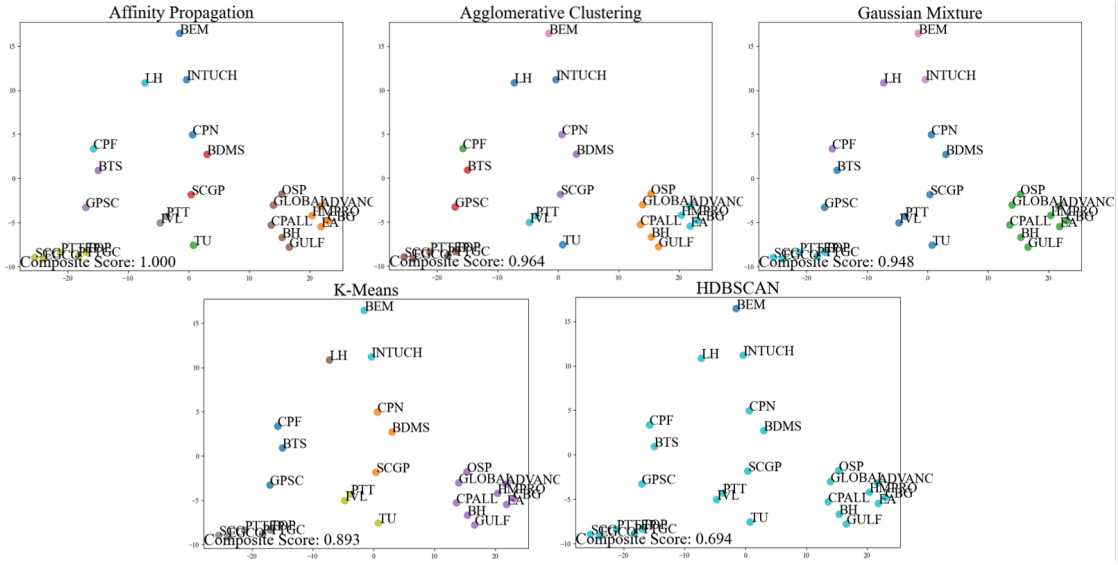

Figure 3 A visualization of clustering results from five algorithms applied to 29 stocks in 2021

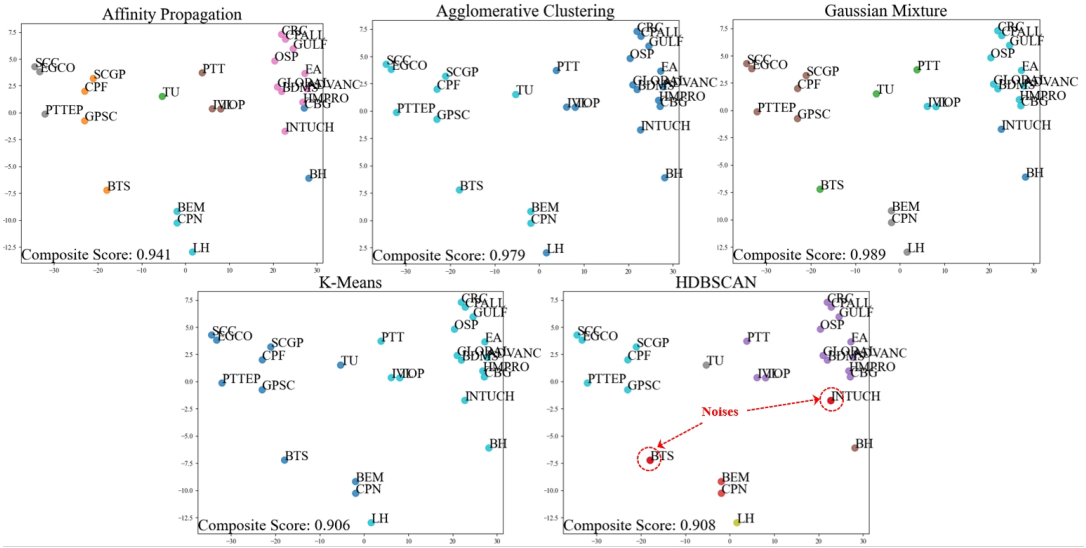

Figure 4 A visualization of clustering results from five algorithms applied to 29 stocks in 2022

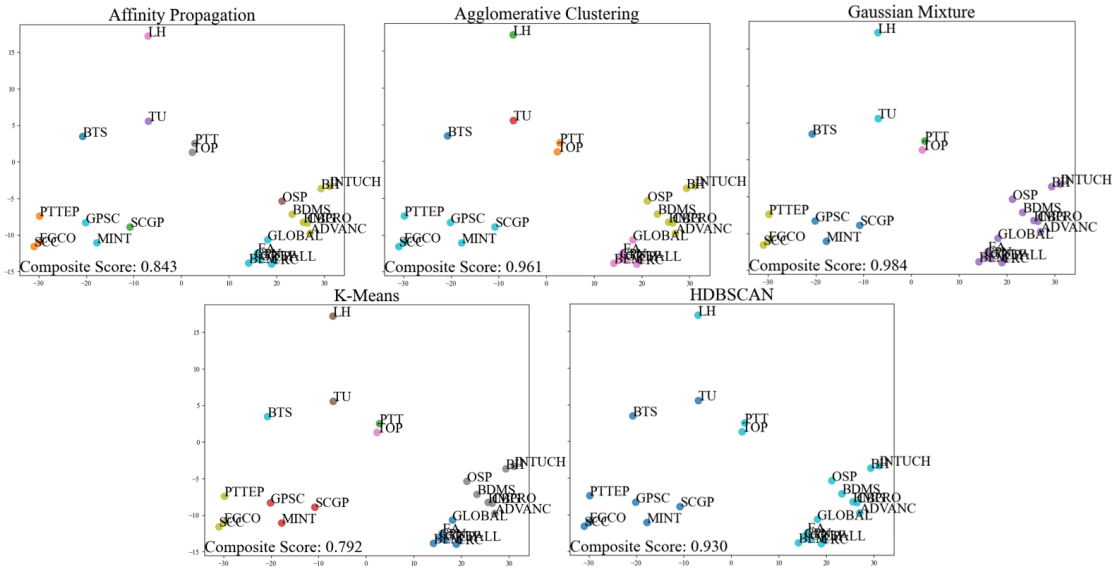

Figure 5 A visualization of clustering results from five algorithms applied to 29 stocks in 2023

Table 4 shows clusters produced by the optimal algorithm from the year 2020–2025, while their graphical views are shown in **Figures 2–5**. **Figure 2** illustrates the clusters of 29 selected stocks in 2020. The cluster produced by the Gaussian Mixture algorithm achieves the highest composite score, as shown in the lower left corner of each graph. Nearly all businesses were affected by the COVID–19 pandemic. The decline in their financial performance resulted in fewer and larger clusters. Gaussian Mixture classified the following set of stocks in the same group: {ADVANC, CBG, HMPRO, OSP, EA, GULF, CPALL, GLOBAL, BDMS}, {CPF, BTS}, {TU, LH}, {BEM}, and {PTT, GPSC, PTTEP, EGCO, SCC}. We discard CPN and INTUCH from consideration since HDBSCAN identified them as noise. Notably, CRC, BH, IVL, MINT, MTC, PTTGC, SCGP, and TOP appear missing from the clustering. This is because these data points are proximate in the high-dimensional space. When reduced to two dimensions using t-SNE for visualization, these points may overlap or be positioned very closely together, giving the impression that some points are missing. Affinity Propagation, Agglomerative, and Gaussian Mixture are the top three algorithms that possess the best composite index.

The year 2021 continued to be impacted by the COVID–19 pandemic; however, most businesses have adapted with various strategies, resulting in more clusters than the previous year. Affinity Propagation attains the best composite index as shown in **Figure 3**. The three largest clusters it created are {ADVANC, HMPRO, CBG, EA}, {EGCO, PTTEP, PTTGC, SCC, TOP}, and {OSP, GLOBAL, CPALL, BH, GULF}. Notably, HDBSCAN detected no stocks as noise or outliers this year. Similar to 2020, the three clustering algorithms with the best composite index are still Affinity Propagation, Agglomerative, and Gaussian Mixture.

Table 4 The group memberships under the optimal clustering algorithm

Year		Clusters
2020	GM	#1 [ADVANC, CBG, HMPRO, OSP, EA, GULF, CPALL, GLOBAL, BDMS]
		#2 [CPF, BTS]
		#3 [TU, LH]
		#4 [BEM]
		#5 [PTT, GPSC, PTTEP, EGCO, SCC]
2021	AP	#1 [ADVANC, HMPRO, CBG, EA]
		#2 [EGCO, PTTEP, PTTGC, SCC, TOP]
		#3 [OSP, GLOBAL, CPALL, BH, GULF]
2022	GM	#1 [CRC, CPALL, GULF, OSP, EA, GLOBAL, BDMS, ADVANC, CBG, HMPRO, IVL, TOP]
		#2 [SCC, EGCO, SCGP, CPF, PTTEP, GPSC]
2023	GM	#1 [BH, INTUCH, OSP, BDMS, CBG, HMPRO, ADVANC, GLOBAL, EA, BEM, CPN, CPALL, CRC, GULF]
		#2 [BTS, PTTEP, EGCO, SCC, MINT, GPSC, SCGP, TU]

The year 2022 marked the beginning of recovery for businesses following the crisis of the previous year. The performance of many companies improved significantly, reflecting their true potential through their financial ratios. In **Figure 4**, most clustering algorithms can form larger clusters than in the previous year. Gaussian Mixture continues to outperform, securing the top composite index. The clusters it produced include {CRC, CPALL, GULF, OSP, EA, GLOBAL, BDMS, ADVANC, CBG, HMPRO, IVL, TOP} and {SCC, EGCO, SCGP, CPF, PTTEP, GPSC}. INTUCH and BTS are discarded since they were identified as noises or outliers by HDBSCAN.

Most companies continued to perform well in 2023, similar to their performance in 2022. This resulted in large clusters of stocks, as shown in **Figure 5**. Gaussian

Mixture still produces clusters with the best composite score. The largest cluster it produced consists of {BH, INTUCH, OSP, BDMS, CBG, HMPRO, ADVANC, GLOBAL, EA, BEM, CPN, CPALL, CRC, GULF}. This grouping may appear unusual when considering the nature of business operations. However, the classification of stocks in this research focuses on the similarity of 14 financial ratios. Therefore, stocks from different sectors may be grouped in the same cluster. This year, businesses exhibited increased stability, resulting in the HDBSCAN algorithm detecting neither noise nor outliers.

To improve the interpretability of the clusters' financial ratio characteristics, **Table 5** summarizes statistics of average key financial ratios (e.g., P/E, ROE, ROA, D/E) for each group.

Table 5 The averages of key financial ratios for each cluster

Year	Cluster ID	P/E	ROE	ROA	D/E
2020	#1	0.0169	0.1907	0.1925	0.0692
	#2	0.0053	0.1479	0.1465	0.0695
	#3	0.0042	0.1224	0.1176	0.0603
	#4	0.0254	0.0496	0.0673	0.0795
	#5	0.0075	0.0702	0.1008	0.0412
2021	#1	0.0344	0.1865	0.1187	0.0215
	#2	0.0094	0.0739	0.0767	0.0146
	#3	0.0582	0.0891	0.0776	0.0144
2022	#1	0.0460	0.1012	0.1653	0.0236
	#2	0.0394	0.0339	0.1053	0.0168
2023	#1	0.0098	0.0863	0.2234	0.0645
	#2	0.0246	0.0316	0.1129	0.0691

To confirm the correctness of the results from the clustering algorithm, we conducted a comparative analysis of stocks within the same cluster and those from different clusters. These financial ratios are rescaled from -1 to +1 to make them visually comparable on the same scale. We aim to cluster stocks primarily using financial ratios. Therefore, stocks in different industries may exhibit similar financial statements reflected in these financial ratios. **Figures 6–9** illustrate the financial ratio comparisons from 2020 to 2023. Due to the space limitation, two closely related stocks from the same cluster and another stock from a different cluster are selected for each comparison.

Figure 6 shows the financial similarity between ADVANC and BH. ADVANC demonstrates the strongest financial performance across most dimensions, particularly in profitability, operational efficiency, and asset utilization. BH maintains a balanced profile with moderate profitability and low leverage, while LH, despite its high dividend yield and efficient fixed asset use, lags in profitability and operational efficiency. Comparing ADVANC and BH, LH is substantially less valued in terms of book value and equity returns, and it operates with significantly lower leverage and asset efficiency. On the other hand, LH stands out with a much higher dividend yield and profitability margins, as well as more active investment and financing cash flows. As a result, LH is isolated from ADVANC and BH. This fact is aligned with the clustering results produced by most algorithms, as LH has never been clustered in the same group as ADVANC and BH.

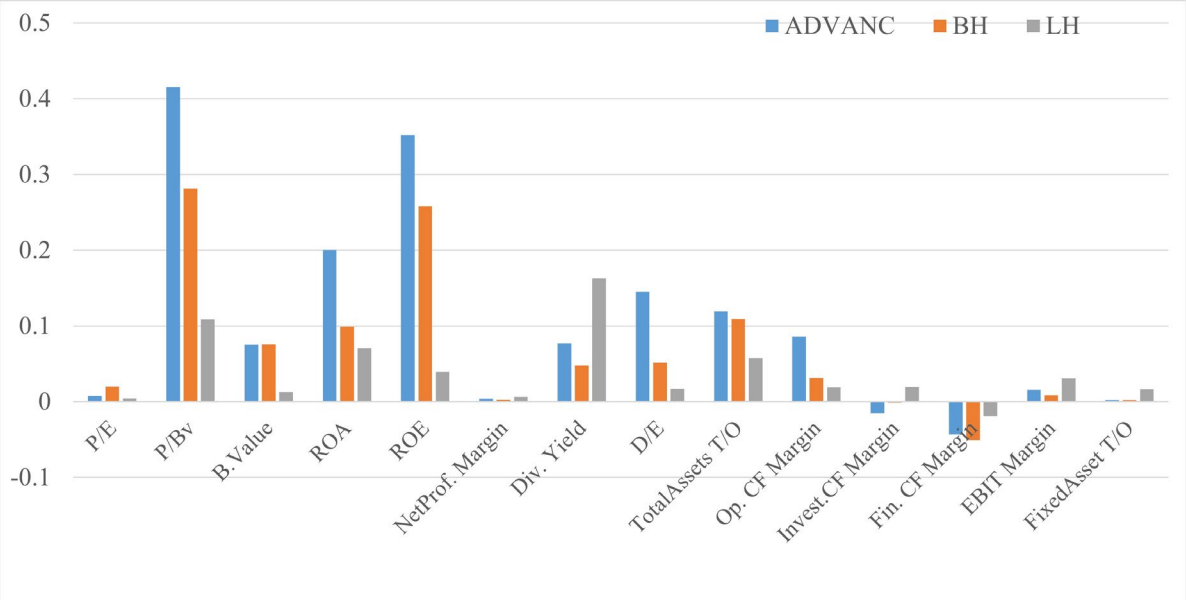

Figure 6 Financial ratio comparisons among ADVANC, BH, and LH in 2020

Comparisons in **Figure 7** indicate that PTT and IVL, which are in the same cluster, exhibit similar financial profiles characterized by moderate profitability, efficient asset utilization, and conservative leverage. IVL edges ahead in shareholder returns and asset turnover. BEM, while less profitable and efficient in asset use overall, distinguishes itself with strong operational cash flow and fixed asset efficiency, likely reflecting the capital-intensive and stable nature of its infrastructure business. These findings highlight the

importance of industry context in interpreting financial ratios and suggest that PTT and IVL are more comparable peers, while BEM operates under a distinct financial model.

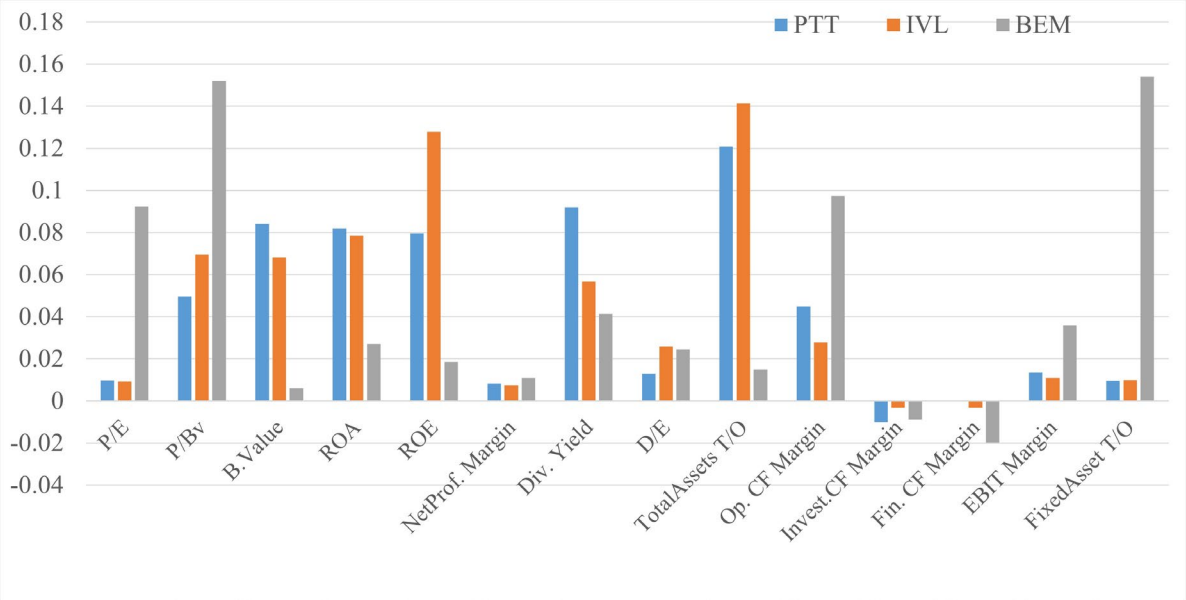

Figure 7 Financial ratio comparisons among PTT, IVL, and BEM in 2021

Figure 8 indicates that HMPRO emerges as the strongest performer overall, with superior profitability, valuation, and operational efficiency. GLOBAL follows closely, offering a balanced profile with moderate valuation and strong efficiency. CPF, while offering the highest dividend yield, underperforms in profitability and cash flow metrics, and its negative investment cash flow suggests a capital-intensive strategy. These findings highlight the strategic and financial distinctions among the three firms, with HMPRO and GLOBAL being more comparable peers in the retail sector, while CPF operates under a different financial model reflective of its agribusiness focus.

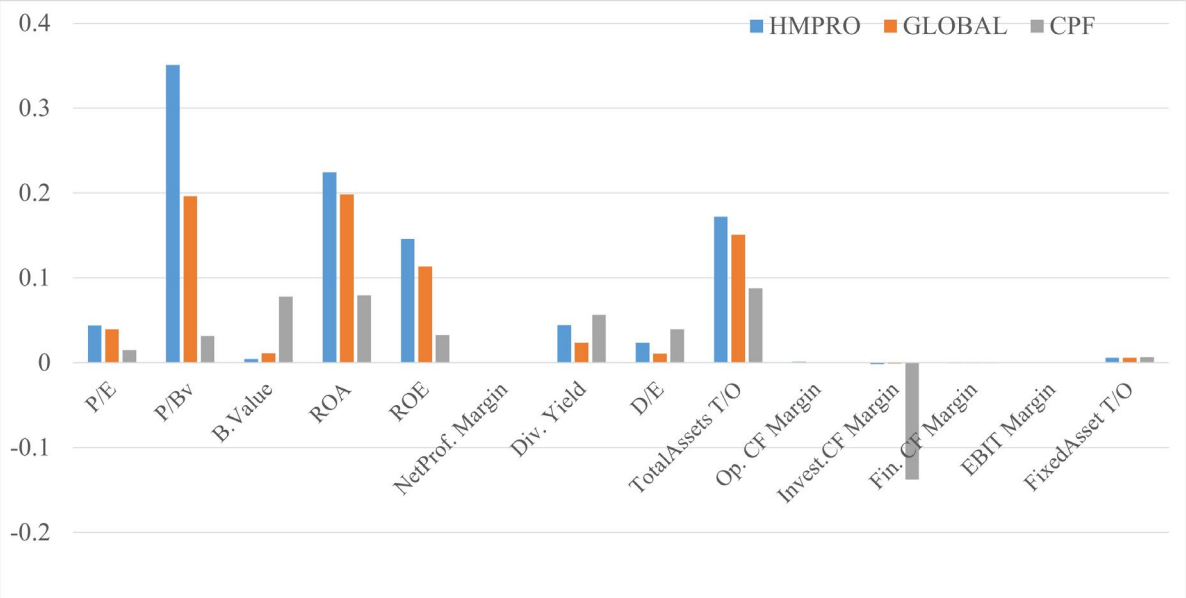

Figure 8 Financial ratio comparisons among HMPRO, GLOBAL, and CPF in 2022

As shown in **Figure 9**, PTTEP stands out as the most profitable and conservatively financed firm, with strong cash flow and dividend performance, albeit with lower asset turnover due to its capital-intensive nature. CPALL demonstrates strong equity efficiency and operational turnover, while GULF balances moderate profitability with positive financing flows. These distinctions reflect the differing industry dynamics and strategic orientations of the three firms, with PTTEP excelling in profitability and shareholder returns, and CPALL and GULF offering more balanced operational profiles.

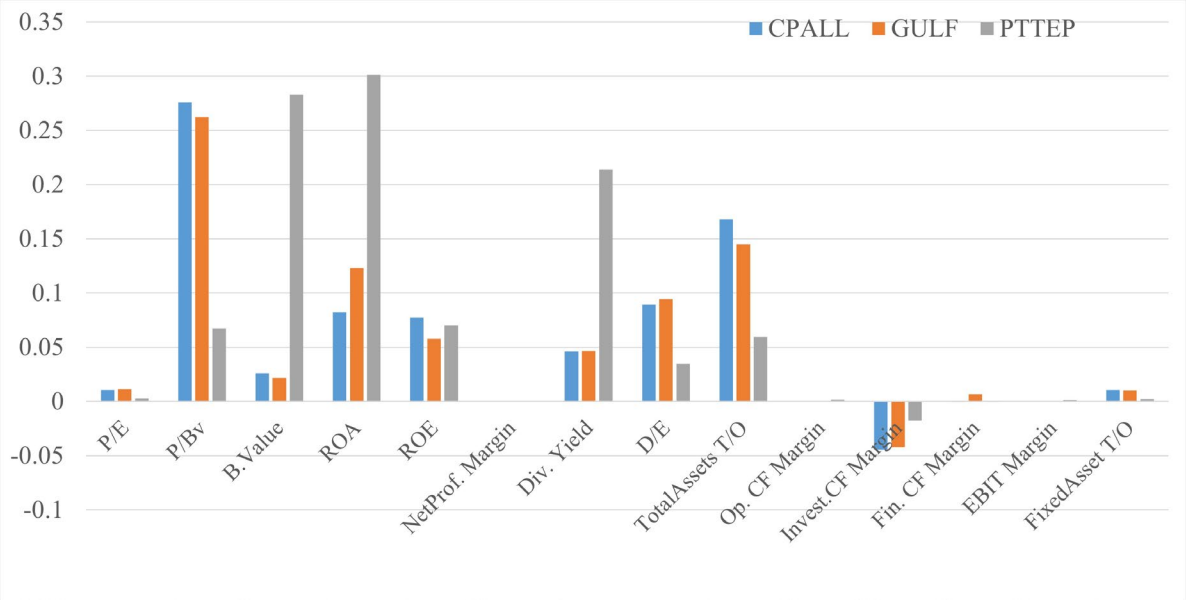

Figure 9 Financial ratio comparisons among CPALL, GULF, and PTTEP in 2023

To conclude the first objective, the clustering algorithm that produces the highest composite index from the past 4-year financial data is the Gaussian mixture model (GMM). GMM is well-suited for continuous, normally distributed data like rescaled financial ratios. It models the data as a mixture of several Gaussian distributions, which can capture the underlying structure even when clusters overlap. For large clusters, GMM can effectively model such scenarios by adjusting the covariance structure of the clusters.

4.4 Result for the second objective: Stocks that are often clustered together based on historical financial ratios

The clustering analysis of historical financial data has yielded insightful groupings of Thai publicly listed companies, revealing underlying structural similarities in their financial and operational characteristics. By applying various clustering algorithms and identifying consensus across methods, three robust and recurrent clusters have emerged, each reflecting distinct corporate profiles and strategic orientations. According to the clustering results from the past financial data, we identify a set of stocks frequently assigned to the same cluster by most clustering algorithms.

The first cluster consists of {ADVANC, CBG, HMPRO}. This cluster comprises firms that are characterized by strong consumer-facing operations, high operational efficiency, and consistent profitability. ADVANC (telecommunications), CBG (consumer beverages), and HMPRO (home improvement retail) share a commonality in their stable cash flows, moderate capital intensity, and relatively high return on assets and equity. These firms typically operate in sectors with steady demand and exhibit resilience to macroeconomic fluctuations. Their inclusion in the same cluster suggests a shared financial structure marked by efficient asset utilization, moderate leverage, and strong market valuation metrics.

The second cluster {CPALL, GULF, GLOBAL, CRC} groups together firms with expansive operational footprints and aggressive growth strategies. CPALL (convenience retail), GULF (energy infrastructure), GLOBAL (construction retail), and CRC (department stores and retail) are unified by their high asset turnover, moderate profitability, and significant investment activities. These companies often engage in capital-intensive expansion, reflected in their investment cash flow patterns and moderate-to-high debt levels. The clustering indicates a strategic orientation toward scale and market penetration, supported by robust revenue generation and diversified operations.

The third cluster {EGCO, PTTEP, SCC} is composed of capital-intensive, asset-heavy firms operating in the energy and industrial sectors. EGCO (electricity generation), PTTEP (oil and gas exploration), and SCC (industrial conglomerate) exhibit high book values, substantial fixed asset bases, and strong operating margins. These firms are typically characterized by lower asset turnover but higher returns on invested capital, reflecting the nature of their long-term infrastructure and resource-based business models. Their clustering underscores a shared emphasis on capital efficiency, long-term investment horizons, and stable dividend policies.

In summary, the clustering results not only validate the financial coherence within each group but also highlight the strategic and sectoral distinctions across the Thai corporate landscape. These insights can inform portfolio diversification strategies, sectoral benchmarking, and risk assessment frameworks. Future research may extend this analysis by incorporating time-series dynamics, macroeconomic variables, or ESG (Environmental, Social, and Governance) indicators to further refine the clustering and enhance its predictive utility.

5. Conclusion

We studied the clustering of stocks in the Stock Exchange of Thailand using 14 key financial ratios. The experiment was conducted with 29 representative stocks from the SET50 index that had complete data for all 14 financial ratios. We found that the stocks frequently clustered together by most algorithms are {ADVANC, CBG, HMPRO}, {CPALL, GULF, GLOBAL, CRC}, and {EGCO, PTTEP, SCC}. Moreover, the Gaussian Mixture algorithm produces the best composite index in almost every year, except for 2021, when Affinity Propagation performed better. The 2023 financial analysis of CPALL, GULF, and PTTEP strongly validates the accuracy of the stock clustering results. A hybrid framework of 5 clustering algorithms provides a comprehensive approach to clustering. Affinity Propagation autonomously determines the number of clusters and handles complex shapes. Agglomerative Clustering builds a hierarchy for deeper insights and offers customizable linkage criteria. Gaussian Mixture provides a probabilistic view and fits diverse cluster shapes. K-Means is simple, fast, and scales well for large datasets, offering distinct clusters. HDBSCAN is highly effective at detecting noise and outliers within a dataset, making it a perfect choice for noise removal. By leveraging their unique strengths, this hybrid approach adapts to various data and enhances clustering accuracy and insight. In addition, the proposed composite index offers a more dependable and detailed evaluation, minimizing the likelihood of skewed outcomes from any single metric and improving the decision-making process for choosing the best clustering algorithm and parameters. We expect the proposed framework to be useful in providing information about stocks with similar financial performance, serving as a guideline for selecting stocks with characteristics similar to those of interest to investors.

6. References

[1] X. Wang, T. Lei, Z. Liu and Z. Wang, "Long-memory behavior analysis of China stock market based on Hurst exponent," in *2017 29th Chinese Control And Decision Conference (CCDC)*, Chongqing, China, May 28–30, 2017, pp. 1710–1712, doi: 10.1109/ccdc.2017.7978792.

[2] K. Jearanaitanakij and B. Passaya, "Predicting Short Trend of Stocks by Using Convolutional Neural Network and Candlestick Patterns," in *2019 4th International Conference on Information Technology (InCIT)*, Bangkok, Thailand, Oct. 24–25, 2019, pp. 159–162, doi: 10.1109/incit.2019.8912115.

[3] F. Liu, X. Li and L. Wang, "Exploring Cluster Stocks based on deep learning for Stock Prediction," in *2019 12th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, Dec. 14–15, 2019, pp. 107–110, doi: 10.1109/iscid.2019.10107.

[4] J. Lee, R. Kim, Y. Koh and J. Kang, "Global Stock Market Prediction Based on Stock Chart Images Using Deep Q-Network," *IEEE Access*, vol. 7, pp. 167260–167277, 2019, doi: 10.1109/access.2019.2953542.

[5] D. Indriyanti and A. Dhini, "Clustering High-Dimensional Stock Data using Data Mining Approach," in *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, Shenzhen, China, Jul. 13–15, 2019, pp. 1–5, doi: 10.1109/icsssm.2019.8887724.

[6] J. H. Moedjahedy, R. Rotikan, W. F. Roshandi and J. Y. Mambu, "Stock Price Forecasting on Telecommunication Sector Companies in Indonesia Stock Exchange Using Machine Learning Algorithms," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, Manado, Indonesia, Oct. 27–28, 2020, pp. 1–4, doi: 10.1109/icoris50180.2020.9320758.

[7] Y. Patil and M. Joshi, "Cluster Driven Candlestick Method for Stock Market Prediction," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, Jul. 3–4, 2020, pp. 1–5, doi: 10.1109/icscan49426.2020.9262356.

[8] Y. Shirota and A. Murakami, "Long-term Time Series Data Clustering of Stock Prices for Portfolio Selection," in *2021 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, Singapore, Dec. 11–12, 2021, pp. 1–6, doi: 10.1109/soli54607.2021.9672407.

[9] X. Wang, K. Yang and T. Liu, "Stock Price Prediction Based on Morphological Similarity Clustering and Hierarchical Temporal Memory," *IEEE Access*, vol. 9, pp. 67241–67248, 2021, doi: 10.1109/access.2021.3077004.

[10] N. Naik and B. R. Mohan, "Novel Stock Crisis Prediction Technique—A Study on Indian Stock Market," *IEEE Access*, vol. 9, pp. 86230–86242, 2021, doi: 10.1109/access.2021.3088999.

[11] T. Leangarun, P. Tangamchit and S. Thajchayapong, "Stock Price Manipulation Detection Using Deep Unsupervised Learning: The Case of Thailand," *IEEE Access*, vol. 9, pp. 106824–106838, 2021, doi: 10.1109/access.2021.3100359.

[12] T. Ploysuwan and N. Pravithana, "Thailand Stock Similarity Clustering by Self-Supervised Wavelet transforms," in *2021 2nd International Conference on Big Data Analytics and Practices (IBDAP)*, Bangkok, Thailand, Aug. 26–27, 2021, pp. 53–57, doi: 10.1109/ibdap52511.2021.9552076.

[13] J. -S. Kim, S. -H. Kim and K. -H. Lee, "Portfolio Management Framework for Autonomous Stock Selection and Allocation," *IEEE Access*, vol. 10, pp. 133815–133827, 2022, doi: 10.1109/access.2022.3231889.

[14] S. Wang, "A Stock Price Prediction Method Based on BiLSTM and Improved Transformer," *IEEE Access*, vol. 11, pp. 104211–104223, 2023, doi: 10.1109/access.2023.3296308.

[15] Y. Li, L. Chen, C. Sun, G. Liu, C. Chen and Y. Zhang, "Accurate Stock Price Forecasting Based on Deep Learning and Hierarchical Frequency Decomposition," *IEEE Access*, vol. 12, pp. 49878–49894, 2024, doi: 10.1109/ACCESS.2024.3384430.

[16] A. Chakravorty and N. Elsayed, "A Comparative Study of Machine Learning Algorithms for Stock Price Prediction Using Insider Trading Data," in *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, Mount Pleasant, MI, USA, Apr. 5–6, 2025, pp. 1–5, doi: 10.1109/icmi65310.2025.11141127.

[17] W. J. Bruns Jr., "Introduction to Financial Ratios and Financial Statement Analysis," Harvard Business School, Boston, MA, USA, Background Note No. 193-029, 1992(Revised Sep. 2004).

[18] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Berkeley Symposium on Mathematical Statistics and Probability*, L. M. Le Cam and J. Neyman, Eds. Berkeley, CA, USA, 1967, pp. 281–297.

[19] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007, doi: 10.1126/science.1136800.

[20] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *The University of Kansas science bulletin*, vol. 38, no. 22, pp. 1409–1438, 1958.

[21] C. E. Rasmussen, "The Infinite Gaussian Mixture Model," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen and K. -R. Müller, Eds. Denver, CO, USA, 1999, pp. 554–560.

[22] R. J. G. B. Campello, D. Moulavi and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining 2013*, J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu, Eds. Gold Coast, Australia, 2013, pp. 160–172.

[23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[24] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[25] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[26] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.