ISSN 2723-6471

Feature Selection Methods as Attribute Weighting Schemes for Clustering Corporate Financial Statements

Tora Fahrudin[1,*], Dedy Rahman Wijaya[2]

[1,2]*School of Applied Science, Telkom University, Bandung, Indonesia*

(Received: February 15, 2026; Revised: April 10, 2026; Accepted: June 18, 2026; Available online: July 12, 2026)

Abstract

Financial clustering plays an important role in uncovering hidden patterns and supporting decision-making in high-dimensional financial statement analysis. However, clustering performance is often constrained by noisy, redundant, and heterogeneous attributes that reduce cluster quality, interpretability, and robustness. Although previous studies in financial analytics have extensively explored clustering algorithms, limited attention has been given to systematically evaluating feature-selection-based attribute weighting strategies for improving clustering effectiveness in complex financial datasets. This study investigates the effectiveness of 6 feature-selection-based weighting strategies for improving financial statement clustering using annual financial data from 604 publicly listed companies on the Indonesia Stock Exchange during 2020-2023. Rather than relying solely on feature filtering, the evaluated methods were utilized to prioritize informative financial attributes and improve clustering structure. Clustering performance was assessed using internal validation metrics, including the Silhouette Coefficient, Dunn Index, Calinski Harabasz Score, and Davies Bouldin Index to evaluate cluster compactness, separation, and overall quality. The results demonstrate that feature-selection-based weighting substantially improves clustering quality compared with the baseline. Fisher Score achieved the strongest overall performance with a Silhouette score of 0.9899, Dunn Index of 1.7238, and Davies Bouldin Index of 0.0034, outperforming the baseline values of 0.8678, 0.3509, and 0.8716, respectively. Autoencoder-based ranking also produced highly competitive results, achieving a Silhouette score of 0.9757 and a Calinski Harabasz Index of 37,105.45. These findings indicate that prioritizing informative financial attributes significantly enhances cluster compactness, separation, and interpretability. The study contributes a comprehensive comparative evaluation of feature-selection-based weighting schemes and provides practical insights for financial segmentation, risk profiling, decision support, and data-driven financial analytics. This study further offers a foundation for developing more robust clustering frameworks for complex financial datasets.

Keywords: Financial Clustering, Feature-Selection-Based Attribute Weighting, Clustering Performance, High-Dimensional Financial Statement Analysis, K-Means, Agglomerative Hierarchical Clustering

1. Introduction

Clustering has been a cornerstone of financial data analysis for decades, serving as a fundamental tool for uncovering hidden patterns and segmenting complex datasets into meaningful groups. The growing volume and complexity of financial data have necessitated the development of advanced techniques for data analysis and clustering. Financial datasets are inherently heterogeneous, often combining transactional histories, credit-related indicators, demographic characteristics, and market-driven variables, each contributing different forms of informational value and uncertainty [1], [2], [3], [4]. Prior studies have demonstrated that transactional and behavioral financial records capture operational patterns [1], credit-related variables reflect risk and solvency characteristics [2], demographic factors influence financial decision-making behavior [3], while market trends introduce external volatility and macroeconomic dynamics [4].

Collectively, these studies highlight that financial data are not only high-dimensional but also exhibit varying levels of feature relevance, redundancy, and interdependence, creating substantial challenges for clustering and pattern discovery. Such complexity motivates the need for advanced analytical techniques, particularly feature selection and weighting mechanisms, to identify informative variables while minimizing noise and redundancy. Advanced techniques for financial data analysis increasingly incorporate machine learning algorithms [5], dimensionality

***Corresponding author: Tora Fahrudin (torafahrudin@telkomuniversity.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i3.1410

reduction approaches [6], and feature selection methods [7] to efficiently process complex financial accounting datasets [8], uncover hidden patterns [9], and group firms into meaningful categories that support interpretability and decision-making [10].

Research in clustering financial datasets has advanced significantly, incorporating various techniques to address specific challenges such as Advanced Clustering Algorithms [11], [12], Integration Clustering with Machine Learning [13], Attribute Weighting Schemes, and Scalability for Big Data. One important challenge in financial clustering is determining how feature selection methods can be utilized as attribute weighting schemes to improve clustering performance. Financial datasets often contain high-dimensional and heterogeneous attributes, where not all features contribute equally to the clustering objective. Irrelevant or redundant features can introduce noise, reduce efficiency, and obscure meaningful patterns. By leveraging feature selection techniques to identify and assign appropriate weights to attributes, clustering algorithms can better reflect the importance of each feature, leading to more accurate and interpretable groupings. This approach not only addresses the challenges of high-dimensionality and feature relevance but also improves the scalability and precision of clustering in financial applications such as customer segmentation [14], fraud detection [15], and credit risk assessment [16]. Therefore, exploring effective feature-weighting strategies is essential for improving clustering robustness and interpretability in increasingly complex financial environments.

The Indonesia Stock Exchange (IDX) provides a challenging setting for financial clustering due to its heterogeneous industrial composition, complex financial reporting structures, and post-pandemic market volatility. This study analyzes financial statements from 604 publicly listed companies across 11 industrial sectors during 2020–2023, comprising 264 financial accounts derived from balance sheets, cash flow statements, and income statements. Such diversity introduces substantial feature interdependence, redundancy, and noise, limiting the effectiveness of conventional clustering approaches. Although financial clustering research has expanded, prior studies have primarily focused on customer segmentation, fraud detection, credit risk assessment, or forecasting, while limited attention has been devoted to clustering corporate financial statements in emerging markets. Moreover, existing studies predominantly emphasize dimensionality reduction or static feature filtering, with limited investigation into feature-selection outputs as attribute-weighting mechanisms for improving clustering robustness and interpretability in high-dimensional financial data.

To address this gap, this study evaluates multiple feature selection methods as attribute-weighting schemes for financial statement clustering while preserving the complete feature space. Rather than eliminating variables, feature importance scores are utilized to prioritize informative financial attributes and reduce the influence of noisy features during clustering. The evaluated approaches include genuine unsupervised feature selection methods (Variance Threshold, Spectral Analysis using Laplacian Score, Multi-Cluster Feature Selection, and Information-Theoretic Approaches) alongside adapted feature-ranking methods (Autoencoder-based ranking and Fisher Score using pseudo-cluster labels from K-Means). Unlike PCA, RFE, or supervised embedded methods that reduce interpretability or require labeled data, the proposed framework prioritizes interpretable feature weighting suitable for unsupervised financial clustering. By integrating statistical, graph-based, clustering-aware, information-theoretic, and deep learning perspectives, this study systematically compares feature-selection-based weighting strategies and identifies Fisher Score and Autoencoder-based ranking as the most effective approaches for high-dimensional financial analytics.

2. Literature Review

Nowadays, applied data science has been implemented in most domains. One of them is in financial data. Previous studies highlighted the importance of clustering in financial data analysis. Research by [17] applied cluster analysis methods to segment companies based on their financial strategies data. Another study by [18] employed cluster and histogram-based anomaly detection to find anomalies in general ledgers of a real company in Bosnia and Herzegovina, demonstrating its effectiveness in identifying outliers. However, both studies noted that the performance of clustering algorithms heavily depends on the quality and relevance of the input features.

Feature selection, which is mentioned as the way to improve the quality and relevance of the input features, has been widely explored to enhance clustering performance in various domains. Mutual information-based methods have been effective in identifying relevant features that effectively select a less redundant and give more relevant set of features

[19]. Principal Component Analysis (PCA) is another popular technique, often used to reduce dimensionality and capture data distribution in datasets [20]. Recursive Feature Elimination (RFE), as demonstrated by [21], has shown promise in reducing uninformative features in nanomaterial grouping classification. Both Laplacian score and K-means for damage characterization in acoustic [22]. Variance Threshold is a simple feature selection technique that removes features with low variance [23]. The underlying assumption is that features with low variance are less informative. This method is often used as a preprocessing step in machine learning workflows to reduce dimensionality and improve model performance such as in Intrusion Detection System study case [24] and Magelang Duck Egg Candling Image [25]. However, this study focuses on unsupervised feature selection approaches that are more suitable for clustering-based financial segmentation

Laplacian Score is a spectral feature selection method that evaluates the quality of features based on the local structure of the data. It aims to preserve the local neighborhood information while reducing dimensionality [26]. Unsupervised spectral feature selection algorithms were applied for omics datasets [27] and gene expression profiles in human cancers [28]. MCFS is another feature selection method that identifies relevant features by leveraging the clustering structure of the data. It selects features that contribute to the formation of multiple clusters [29]. Autoencoders are neural networks used for unsupervised learning of efficient coding's. They can also be employed for feature ranking by analyzing the reconstruction error or the learned representations [30]. The Fisher Score is a supervised feature selection method that evaluates features based on the ratio of between-class variance to within-class variance [31]. It is widely used in classification tasks. Information-theoretic approaches for feature selection focus on maximizing the information gained from features concerning the target variable. An example of this method is for high dimensional data [32].

Several researchers have incorporated attribute weighting schemes to improve clustering results. For example, [33] proposed a weighted K-Means algorithm for groundwater resource planning that assigns weighted based on radius and mean value. Similarly, a study by [34] integrated AHP-Entropy weight method and the improved Euclidean distance calculation to obtain more accurate clustering results in segmenting e-commerce customer method based on RFM Model. Despite these advances, there is a need for further exploration of weighting schemes tailored to financial datasets [35], particularly those integrating real-world complexities such as missing values [36], handling mix type of dataset [37], transfer relationships of attribute weighting between groups [38], and noise [39].

Evaluating the effectiveness of attribute weighting schemes in clustering tasks involves several key metrics that assess both the quality of the clustering and the appropriateness of the assigned weights. Commonly used metrics include Precision [36], [40], [41], [42], Recall [36], [38], [40], [42], Accuracy [37], [41], [42], [43], [44], [45]. Additionally, the F1-score is utilized by [36], [46] to evaluate weighting performance by comparing the sets of features with the highest weights, providing insight into the relevance of selected attributes. These metrics collectively offer a comprehensive evaluation framework, ensuring that the attribute weighting schemes enhance clustering performance by accurately reflecting feature importance and improving the overall quality of the clustering results. But another clustering evaluation metrics such as Adjusted Rand index [37], [43], [45], Normalized Mutual Information [47], Silhouette Score [44], Dunn Index, and Davies-Bouldin Index are widely used to evaluate the effectiveness of clustering and can also assess the impact of attribute weighting schemes in clustering tasks.

Although prior studies have shown that attribute weighting can improve clustering performance, most approaches rely on manually designed weighting strategies, domain-specific heuristics, or optimization frameworks, while limited attention has been given to feature-selection-based weighting for unsupervised financial statement clustering. Existing studies also tend to focus on single techniques or non-financial contexts, providing limited comparative understanding of different feature-prioritization strategies in high-dimensional financial environments. To address this gap, this study systematically compares six feature-selection-based weighting approaches for clustering financial statements of Indonesia Stock Exchange (IDX) companies, including genuine unsupervised feature selection methods (Variance Threshold, Spectral Analysis, Multi-Cluster Feature Selection, and Information-Theoretic Approaches) and adapted feature-ranking approaches (Autoencoder-based ranking and Fisher Score using pseudo-cluster labels). Unlike conventional weighting methods that depend on external optimization or feature elimination, this study preserves the complete feature space and operationalizes feature importance scores as attribute-weighting mechanisms to improve clustering robustness and interpretability across multiple clustering algorithms and validation metrics.

 ISSN 2723-6471

3. Methodology

This section describes the methodology employed to evaluate the effect of feature-selection-based weighting on clustering performance in financial statement data from companies listed on the Indonesia Stock Exchange (IDX). The overall process consists of five main stages: data preprocessing, feature relevance estimation, feature weighting and weighted representation, clustering, and clustering evaluation. To ensure fair distance computation, all numerical features were normalized using Min-Max scaling. Subsequently, six feature-prioritization approaches were applied to estimate attribute importance, including genuine unsupervised feature selection methods such as Variance Threshold, Spectral Analysis (Laplacian Score), Multi-Cluster Feature Selection (MCFS), and Information-Theoretic Approaches, alongside adapted feature-ranking approaches including Autoencoder-based ranking and Fisher Score using pseudo-cluster labels generated through K-Means. Rather than explicitly removing variables, the importance scores generated by each method were transformed into normalized feature weights and applied to the complete feature space to emphasize informative financial attributes while reducing the influence of noisy or less relevant variables. To comprehensively evaluate the impact of feature-selection-based weighting on clustering performance, two clustering algorithms representing different paradigms were employed: K-Means (partition-based) and Agglomerative Hierarchical Clustering (hierarchical-based). Clustering quality was subsequently assessed using multiple internal validation metrics, including the Silhouette Coefficient, Dunn Index, Calinski Harabasz Score, and Davies Bouldin Index. Finally, the proposed weighting strategies were compared against baseline clustering without feature weighting to evaluate their effectiveness in improving cluster compactness, separation, and overall clustering quality. Figure 1 illustrates the proposed methodological framework.

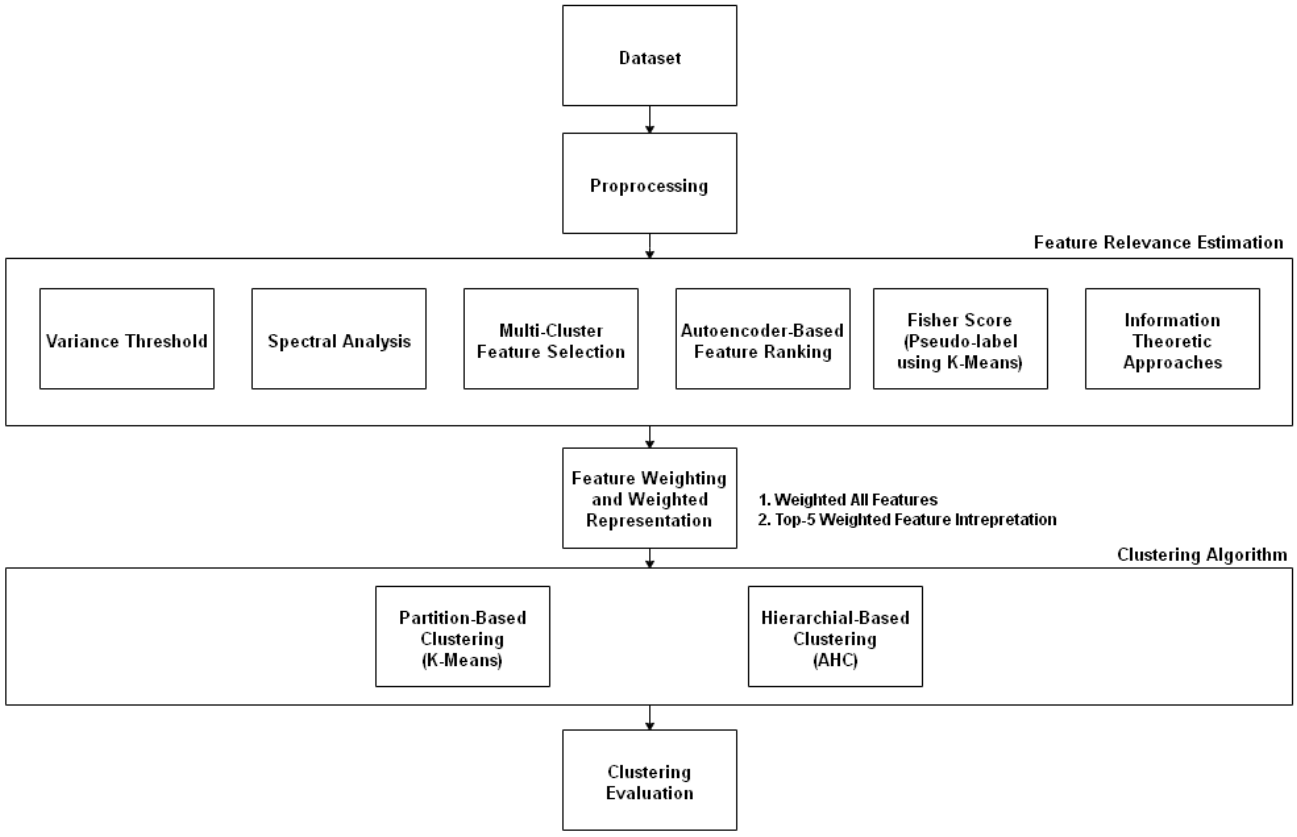

Figure 1. Proposed Method

3.1. Dataset and Computing Infrastructure

This study utilizes a publicly available dataset obtained from Kaggle, consisting of annual financial statement data from 604 publicly listed companies on the Indonesia Stock Exchange (IDX) during 2020–2023. After transforming financial records into a company-feature matrix, the dataset comprised 604 observations and 1,056 financial attributes

derived from 264 financial statement accounts across Balance Sheet (BS), Cash Flow (CF), and Income Statement (IS) categories.

The dataset covers 11 industrial sectors, including Healthcare, Financials, Energy, Technology, Consumer sectors, Transportation & Logistics, and Properties & Real Estate, reflecting substantial heterogeneity in financial reporting structures, capital intensity, and operational characteristics. Although sectoral representation is imbalanced, the original composition was intentionally preserved to maintain the real-world structure of the Indonesian capital market and evaluate clustering performance under authentic financial conditions. All analyses were conducted using Python 3.9 on a Windows 11 (64-bit) system with an Intel Core i7-9750H CPU and 8 GB RAM.

3.2. Preprocessing

The preprocessing stage involves several essential procedures to ensure data quality, consistency, and suitability for feature selection and clustering analysis. Initially, the raw dataset containing financial statement information from 264 accounts was transformed from a row-based representation into a structured columnar format, where each feature corresponds to a specific financial account. These features were categorized into three principal financial statement groups, namely Balance Sheet (BS), Cash Flow (CF), and Income Statement (IS).

The preprocessing stage was conducted to ensure data consistency and suitability for feature weighting and clustering analysis. Initially, financial statement records were transformed into a structured feature matrix where each attribute corresponds to a specific financial account. Exploratory assessment revealed 315,897 missing values (49.48% of the dataset), primarily reflecting heterogeneous reporting practices and non-applicable financial accounts across firms and sectors. To preserve dataset completeness and avoid information loss, missing values were handled using zero imputation (fillna(0)), assuming that unavailable accounts frequently indicate absent or non-reported financial activities rather than recording errors.

To ensure numerical consistency, formatting symbols and comma delimiters were removed prior to analysis. Since financial variables exhibit highly heterogeneous magnitudes, Min-Max normalization was applied to transform features into a standardized range of [0,1], preventing large-scale variables from disproportionately influencing distance-based clustering methods such as K-Means and Agglomerative Hierarchical Clustering (AHC). Explicit outlier removal was not performed, as extreme financial values may represent genuine economic conditions, including financial distress, extraordinary gains, or sector-specific capital structures. Instead, normalization was employed to reduce scale dominance while preserving authentic financial variability across firms. Table 1 shows an example of the dataset.

Table 1. Dataset

symbol	account	type	2020	2021	2022	2023	category
AALI	Accounts Payable	BS	770,264,000,000	1,026,717,000,000	1,224,423,000,000	842,064,000,000	CONSUMER NON CYCLICALS
AALI	Accounts Receivable	BS	765,849,000,000	458,135,000,000	848,770,000,000	674,487,000,000	CONSUMER NON CYCLICALS
AALI	Accumulated Depreciation	BS	(10,920,948,000,000)	(12,133,813,000,000)	(13,303,749,000,000)	(14,436,847,000,000)	CONSUMER NON CYCLICALS
AALI	Additional Paid In Capital	BS	3,878,995,000,000	3,878,995,000,000	3,878,995,000,000	3,878,995,000,000	CONSUMER NON CYCLICALS
AALI	Allowance For Doubtful Accounts Receivable	BS	(24,261,000,000)	(24,543,000,000)	(27,057,000,000)	(26,516,000,000)	CONSUMER NON CYCLICALS
AALI	Basic Average Shares	IS	1,924,688,333	1,924,688,333	1,924,688,333	1,924,688,333	CONSUMER NON CYCLICALS
AALI	Basic EPS	IS	433	1,024	897	549	CONSUMER NON CYCLICALS
AALI	Beginning Cash Position	CF	383,366,000,000	978,892,000,000	3,896,022,000,000	1,619,616,000,000	CONSUMER NON CYCLICALS
AALI	Buildings And Improvements	BS	4,775,744,000,000	4,858,140,000,000	4,926,098,000,000	5,028,178,000,000	CONSUMER NON CYCLICALS
AALI	Capital Expenditure	CF	(999,198,000,000)	(1,229,482,000,000)	(1,379,444,000,000)	(1,236,337,000,000)	CONSUMER NON CYCLICALS

symbol	account	type	2020	2021	2022	2023	category
…	….	…	…	…	…	…	…
ZINC	Total Tax Payable	BS	2,191,522,069	2,123,986,973	1,142,167,582	3,629,701,805	BASIC MATERIALS
ZINC	Total Unusual Items	IS	2,459,340,945	10,966,950,897	747,497,939	81,174,742	BASIC MATERIALS
ZINC	Total Unusual Items Excluding Goodwill	IS	2,459,340,945	10,966,950,897	747,497,939	81,174,742	BASIC MATERIALS
ZINC	Tradeand Other Payables Non Current	BS	154,245,597,129	3,418,800,000	3,418,800,000	3,418,800,000	BASIC MATERIALS
ZINC	Work In Process	BS	375,770,262	375,770,262	3,874,521,715	70,672,919,998	BASIC MATERIALS
ZINC	Working Capital	BS	53,655,888,672	459,909,106,336	(30,236,329,814)	128,536,743,315	BASIC MATERIALS

3.3. Feature Relevance Estimation

This study distinguishes feature selection from attribute weighting, as both concepts are closely related but conceptually different. Conventional feature selection typically aims to identify and remove irrelevant variables to reduce dimensionality, whereas attribute weighting assigns different levels of importance to features during model construction. Although the evaluated methods originate from the feature selection literature, this study operationalizes them as feature-selection-based weighting mechanisms by transforming feature importance scores into normalized attribute weights prior to clustering. Consequently, the complete financial feature space (1,056 attributes) is preserved, while the influence of informative variables is emphasized and noisy or less relevant attributes are attenuated without explicit feature elimination. For example, highly weighted financial indicators such as liquidity, cash position, or investment-related variables exert greater influence during distance computation, thereby improving cluster separability and interpretability in heterogeneous financial data.

To estimate feature relevance, six complementary approaches were employed, consisting of genuine unsupervised feature selection methods and adapted feature-ranking techniques. Variance Threshold was included as a statistical baseline to identify low-information features based on variance. Spectral Analysis using Laplacian Score [48] captures local neighborhood structures and nonlinear relationships among financially similar firms. Multi-Cluster Feature Selection (MCFS) evaluates feature relevance by preserving latent multi-cluster structures through spectral embedding. Information-Theoretic Approaches assess attribute importance using entropy-based information content without relying on labeled classes. In addition, Autoencoder-based ranking was adopted as a deep learning-driven feature-ranking approach capable of modeling nonlinear dependencies in financial statements through reconstruction behavior. Fisher Score was adapted for unsupervised settings by utilizing pseudo-cluster labels generated through K-Means to evaluate inter-cluster separability and intra-cluster compactness. Together, these approaches enable a systematic comparison of statistical, graph-based, clustering-aware, information-theoretic, and deep learning-driven weighting strategies for high-dimensional financial clustering.

3.4.Feature Weighting and Weighted Representation

In unsupervised learning scenarios, particularly for clustering tasks, it is essential to identify and emphasize features that contribute meaningfully to the intrinsic structure of the data. Feature weighting refers to the process of assigning important scores to individual features based on their ability to preserve data topology, local neighborhood structure, or statistical variability.

$$X_{weighted} = X_{normalized} \times norm\_weights \quad (1)$$

These normalized weights in Eq. (1) were subsequently used to reweight the original features by element-wise multiplication, resulting in a feature-weighted dataset. These transformations not only reduce computational complexity but also enhance cluster separability by eliminating redundancy and highlighting hidden data patterns. When used together or in sequence, feature weighting and transformation form a powerful preprocessing strategy to support effective unsupervised clustering and data exploration.

3.5. Clustering

To evaluate the effectiveness of feature-selection-based weighting on clustering quality, two clustering algorithms representing different paradigms were employed: K-Means (partition-based) and Agglomerative Hierarchical Clustering (AHC, hierarchical-based). Prior to the main experiments, an exploratory procedure was conducted to determine an appropriate number of clusters by evaluating candidate configurations ranging from (k = 2) to (k = 5) using multiple internal validation metrics, including the Silhouette Score, Dunn Index, Calinski Harabasz Index, and Davies Bouldin Index. As shown in table 2, although (k = 2) achieved slightly stronger compactness-related performance, (k = 3) was selected as a more balanced and interpretable clustering configuration. In the context of heterogeneous financial statement data from IDX-listed companies, the three-cluster structure provided richer differentiation of firms with varying liquidity, solvency, and operational characteristics while maintaining strong clustering quality. The literature on K-Means clustering can handle categorical and mixed data types [49]. Therefore, (k = 3) was consistently applied across all experiments to ensure methodological consistency and interpretability.

The clustering experiments employed K-Means and AHC to evaluate the impact of feature-selection-based weighting schemes. For K-Means, clustering was implemented using the k-means++ initialization strategy with a fixed 'random_state = 0' to improve centroid stability and ensure reproducibility. Default convergence parameters from the Scikit-learn implementation were retained. For AHC, Ward linkage with Euclidean distance was employed to iteratively merge observations while minimizing within-cluster variance. This linkage criterion was selected due to its ability to generate compact and homogeneous clusters, which is particularly suitable for heterogeneous financial statement data characterized by interdependent accounting attributes. Since feature weights modify the contribution of financial variables during distance computation, the use of both partition-based and hierarchical clustering paradigms enables a more robust assessment of clustering behavior under different structural assumptions. Recent studies have demonstrated that integrating feature weighting into AHC can significantly enhance clustering performance. For instance, algorithm [50] introduces cluster-dependent feature weights using the Lp norm, allowing each feature to have varying relevance across different clusters. This approach effectively resizes features based on their importance, leading to more accurate and meaningful cluster formations. Furthermore, the Weighted Hierarchical Agglomerative Clustering Ensemble (WHAC) method [51] proposes a novel similarity measurement at the cluster level, determining the merit of individual clustering methods and enhancing the robustness of the clustering process.

Table 2. Cluster Number Evaluation Prior to Clustering Experiments

n_cluster	Silhouette	Dunn Index	CalinskiHarabasz	DaviesBouldin
2	0.8742	0.2277	172.6818	1.1432
3	0.8682	0.2277	142.3561	1.1329
4	0.8165	0.1514	130.5335	1.0700
5	0.7786	0.1260	136.4099	1.0096

3.6. Clustering Evaluation

By applying these three clustering algorithms to both the original and the feature-weighted datasets, we assess whether feature importance learned through unsupervised methods improves cluster formation. Clustering performance is evaluated using multiple internal validation metrics such as Silhouette Coefficient, Dunn Index, Calinski-Harabasz Score, and Davies-Bouldin Index, allowing for a multi-faceted analysis of compactness, separability, and cluster quality. To ensure a comprehensive evaluation, each clustering algorithm was applied consistently across datasets transformed by different unsupervised feature selection techniques. This design enables a direct comparison of how each feature selection strategy influences the structural quality of clusters.

4. Results and Discussion

This section evaluates the effectiveness of feature-selection-based weighting schemes in improving clustering quality. Six weighting approaches Variance Threshold, Spectral Analysis (Laplacian Score), Multi-Cluster Feature Selection (MCFS), Autoencoder-based ranking, Fisher Score, and Information-Theoretic Approaches were integrated with K-

ISSN 2723-6471

Means and Agglomerative Hierarchical Clustering (AHC), alongside a baseline without weighting. Clustering quality was assessed using the Silhouette Score, Calinski Harabasz Index, Davies Bouldin Index, and Dunn Index. As shown in table 3, feature-selection-based weighting generally improved clustering quality compared to the baseline, confirming the importance of prioritizing informative financial attributes in high-dimensional financial statement data.

Table 3. Experiment Result of All Features

FeatureSelector	Clustering	Silhouette	CalinskiHarabasz	DaviesBouldin	Dunn
VarianceThreshold	AHC	0.873891912	143.4316737	0.840238372	0.315752072
VarianceThreshold	KMeans	0.883881621	141.7812501	0.825387232	0.358167763
SpectralAnalysis	AHC	0.910242551	1841.826424	**0**	0.285017713
SpectralAnalysis	KMeans	0.910242551	1841.826424	**0**	0.285017713
MCFS	AHC	0.86117588	86.94675505	1.139868732	0.518016933
MCFS	KMeans	0.803374922	88.09732936	1.569478925	0.20255736
Autoencoder	AHC	0.97567325	**37105.44955**	0.009007641	0.631210287
Autoencoder	KMeans	0.97567325	**37105.44955**	0.009007641	0.631210287
FisherScore	AHC	**0.989877415**	3818.588865	0.003436437	**1.723789063**
FisherScore	KMeans	**0.989877415**	3818.588865	0.003436437	**1.723789063**
InformationTheoreticApproaches	AHC	0.862758651	136.7629318	0.854045133	0.290897025
InformationTheoreticApproaches	KMeans	0.861041195	135.8203209	0.531354129	0.326755613
BaseLine	AHC	0.867786887	146.3307913	0.871579585	0.350889338
BaseLine	KMeans	0.867786887	146.3307913	0.871579585	0.350889338

Among the evaluated methods, Fisher Score achieved the strongest and most consistent performance, obtaining the highest Silhouette and Dunn Index values alongside a very low Davies Bouldin Index, indicating highly compact and well-separated clusters. This performance may be attributed to Fisher Score's ability to emphasize features exhibiting strong inter-cluster discrimination while minimizing intra-cluster variation, thereby improving the separation of firms with heterogeneous financial profiles. Similarly, Autoencoder-based ranking demonstrated strong performance, particularly in the Calinski Harabasz Index, reflecting its effectiveness in capturing nonlinear dependencies and latent financial structures embedded in complex accounting relationships. Since financial statements often contain interdependent variables, deep representation learning enables more informative weighting of financially relevant attributes.

Spectral Analysis (Laplacian Score) also produced robust clustering results by preserving local neighborhood relationships among firms, indicating that graph-based weighting can effectively capture structural similarity in financial statements. In contrast, MCFS and Information-Theoretic Approaches showed relatively weaker performance, potentially due to sensitivity to sectoral heterogeneity, sparse financial characteristics, and noisy accounting variables common in IDX financial data.

The findings further suggest that effective feature weighting reduces dependence on clustering algorithms. For stronger weighting approaches such as Fisher Score and Autoencoder-based ranking, K-Means and AHC produced nearly identical results, indicating that informative feature prioritization creates a more stable and separable clustering structure. Conversely, weaker weighting schemes exhibited greater sensitivity to clustering strategy, with AHC generally showing higher robustness in handling heterogeneous financial profiles.

The visualization of clustering results across various methods, as illustrated in the PCA scatter plots in figure 2, demonstrates that incorporating feature selection significantly enhances cluster separability compared to the baseline. While the baseline (figure 2) exhibits dense data accumulation with poorly defined margins, the application of feature

ISSN 2723-6471

weighting techniques such as Fisher Score, Mutual Information, and MCFS effectively suppresses noise and emphasizes discriminative attributes, leading to clearer boundaries along the principal components.

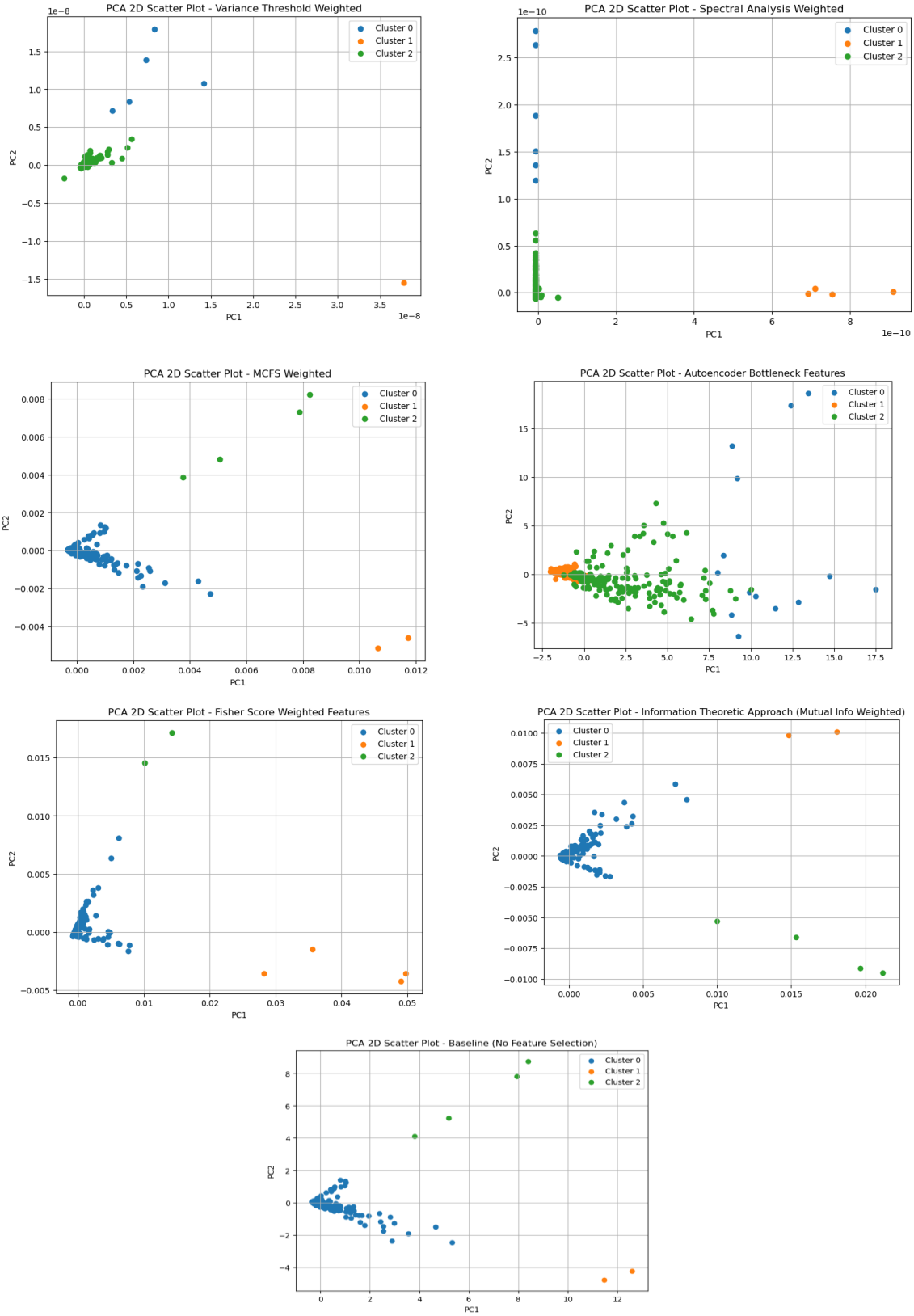

Figure 2. Scatter Plot of Variance; Spectral; MCFS; Autoencoder; Fisher; Information Theoretic; Baseline.

Specifically, Spectral Analysis and Autoencoder-based extraction succeeded in capturing local manifold structures and non-linear relationships, respectively, although some overlap persisted between Cluster 0 and Cluster 2. In contrast,

 ISSN 2723-6471

Information Theoretic Approaches and MCFS provided the most distinct isolation of smaller clusters (Clusters 1 and 2), confirming that strategic feature selection is essential for uncovering latent structures and improving the robustness of clustering outcomes in high-dimensional datasets.

Table 4 presents the top five weighted financial attributes identified by each feature-selection-based weighting method, revealing substantial variation in how financial relevance is prioritized. Variance Threshold emphasized liability-related variables, such as Payables and Interest Expense, indicating that firms are differentiated primarily through financing obligations and debt structures. Spectral Analysis (Laplacian Score) prioritized liquidity and investment-related indicators, including Changes in Cash and Operating Cash Flow, suggesting that firms with similar cash dynamics and investment behavior tend to form structurally similar clusters. MCFS highlighted operational and profitability-related variables, such as Interest Paid CFO, Receivables Changes, and EBIT, reflecting multidimensional financial behavior, although its clustering performance remained comparatively weaker due to greater heterogeneity in selected attributes.

Table 4. Top 5 Features

Methods	1st Column	2nd Column	3rd Column	4th Column	5th Column
VarianceThreshold	2020_Payables	2023_Interest Expense Non Operating	2022_Total Non Current Liabilities Net Minority Interest	2022_Interest Expense Non Operating	2022_Prepaid Assets
SpectralAnalysis	2020_Changes In Cash	2023_Operating Cash Flow	2022_Net Business Purchase And Sale	2022_Net Investment Purchase And Sale	2022_Total Unusual Items Excluding Goodwill
MCFS	2020_Interest Paid Cfo	2020_Change In Receivables	2022_Dividends Paid Direct	2023_Taxes Refund Paid Direct	2023_EBIT
Autoencoder	2023_Long Term Debt Payments	2020_Other Cash Adjustment Inside Changein Cash	2023_Derivative Product Liabilities	2022_Goodwill	2023_Investments In Other Ventures Under Equity Method
FisherScore	2023_Available For Sale Securities	2020_Cash And Cash Equivalents	2021_End Cash Position	2021_Cash And Cash Equivalents	2022_Beginning Cash Position
InformationTheoreticApproaches	2021_Net Non Operating Interest Income Expense	2022_Normalized Income	2023_Operating Income	2023_Net Income Continuous Operations	2023_Net Income Including Noncontrolling Interests

In contrast, Autoencoder-based feature ranking prioritized complex indicators such as Long-Term Debt Payments, Derivative Product Liabilities, and Goodwill, indicating its ability to capture latent nonlinear relationships among solvency, investment, and financing behavior embedded in financial statements. Similarly, Fisher Score consistently emphasized liquidity-related variables, including Cash and Cash Equivalents and Available-for-Sale Securities, suggesting that liquidity management represents a strong discriminative dimension among IDX firms. This behavior helps explain why Fisher Score achieved superior clustering performance, as liquidity indicators exhibit strong between-group discrimination while remaining relatively stable within clusters.

Meanwhile, Information-Theoretic Approaches focused on profitability-related variables such as Operating Income and Net Income, indicating that entropy-based weighting captures informational relevance in earnings behavior. However, maximizing information content alone did not necessarily translate into stronger cluster separability in high-dimensional financial data. Although clustering quality was primarily evaluated using internal validation metrics due to the absence of ground-truth labels, practical interpretability was strengthened through examination of the highly weighted financial indicators. The selected attributes reflect meaningful financial dimensions, including liquidity, solvency, profitability, and investment behavior, supporting substantive interpretation of the resulting firm clusters beyond purely statistical evaluation.

5. Conclusion

This study investigated whether feature selection methods, when operationalized as attribute weighting schemes, can improve clustering performance in financial statement analysis. The findings indicate that feature-selection-based weighting substantially enhances clustering quality by prioritizing informative financial attributes while reducing the influence of noisy variables without eliminating the original feature space. Among the evaluated methods, Fisher Score

and Autoencoder-based feature ranking consistently achieved the strongest performance across clustering validity metrics, demonstrating robustness in handling high-dimensional financial data.

From a practical perspective, the resulting clusters may support investment screening, financial risk profiling, peer benchmarking, and regulatory monitoring by identifying firms with similar liquidity, solvency, profitability, and operational characteristics. These findings are particularly relevant for heterogeneous financial environments such as the Indonesia Stock Exchange (IDX), where firms within the same sector may exhibit substantially different financial profiles.

Several limitations should be acknowledged. The study focuses exclusively on IDX-listed firms during 2020–2023, limiting direct generalizability and introducing sensitivity to macroeconomic conditions, sectoral imbalance, preprocessing choices, and clustering algorithm assumptions. In addition, clustering evaluation relied primarily on internal validity metrics due to the absence of ground-truth labels, while statistical significance testing and computational efficiency analysis were beyond the scope of this study. Future research may strengthen robustness and practical applicability through cross-market datasets, external financial benchmarks, statistical validation, runtime comparisons, and hybrid approaches integrating feature weighting with selective dimensionality reduction.

6. Declarations

6.1. Author Contributions

Conceptualization: T.F. and D.R.W.; Methodology: D.R.W.; Software: T.F.; Validation: T.F. and D.R.W.; Formal Analysis: T.F. and D.R.W.; Investigation: T.F.; Resources: D.R.W.; Data Curation: D.R.W.; Writing Original Draft Preparation: T.F. and D.R.W.; Writing Review and Editing: D.R.W. and T.F.; Visualization: T.F.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This publication funding was supported by the Directorate of Research and Community Service at Telkom University.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Lauer, “Plastic surveillance: Payment cards and the history of transactional data, 1888 to present,” *Big Data Soc.*, vol. 7, no. 1, pp. 81–93, Jan. 2020, doi: 10.1177/2053951720907632.

[2] A. Markov, Z. Seleznyova, and V. Lapshin, “Credit scoring methods: Latest trends and points to consider”, *Journal of Finance and Data Science*, vol. 8, no. 4, pp. 180–201, Nov. 01, 2022, doi: 10.1016/j.jfds.2022.07.002.

[3] H. Kaur and S. Arora, “Demographic influences on consumer decisions in the banking sector: evidence from India,” *Journal of Financial Services Marketing*, vol. 24, no. 3–4, pp. 81–93, Dec. 2019, doi: 10.1057/s41264-019-00067-4.

[4] V. Chang, Q. A. Xu, A. Chidozie, and H. Wang, “Predicting Economic Trends and Stock Market Prices with Deep Learning and Advanced Machine Learning Techniques,” *Electronics (Switzerland)*, vol. 13, no. 17, pp. 1–27, Sep. 2024, doi: 10.3390/electronics13173396.

[5] X. Liu, S. Salem, L. Bian, J. T. Seong, and H. M. Alshanbari, "Application of machine learning algorithms in the domain of financial engineering," *Alexandria Engineering Journal*, vol. 95, no. May, pp. 94–100, May 2024, doi: 10.1016/j.aej.2024.03.058.

[6] D. G. Cortés, E. Onieva, I. P. López, L. Trinchera, and J. Wu, "Autoencoder-Enhanced Clustering: A Dimensionality Reduction Approach to Financial Time Series," *IEEE Access*, vol. 12, no. January, pp. 16999–17009, 2024, doi: 10.1109/ACCESS.2024.3359413.

[7] K. Mialkowska, K. Kaczmarczyk, M. Hernes, and M. Dyvak, "Feature Selection for financial data - comparison," *Procedia Computer Science*, vol. 207, pp. 3041–3050, 2022, doi: 10.1016/j.procs.2022.09.362.

[8] P. Chakri, S. Pratap, Lakshay, and S. K. Gouda, "An exploratory data analysis approach for analyzing financial accounting data using machine learning," *Decision Analytics Journal*, vol. 7, no. June, Art. no. 100212, 2023, doi: 10.1016/j.dajour.2023.100212.

[9] D. Zhang and L. Zhou, "Discovering Golden Nuggets: Data Mining in Financial Application," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 4, pp. 513–522, Nov. 2004, doi: 10.1109/TSMCC.2004.829279.

[10] H. Abdollahi, J. P. Junttila, and H. Lehkonen, "Clustering asset markets based on volatility connectedness to political news," *Journal of International Financial Markets, Institutions and Money*, vol. 93, no. June, Art. no. 102004, 2024, doi: 10.1016/j.intfin.2024.102004.

[11] T. Li, G. Kou, Y. Peng, and P. S. Yu, "An Integrated Cluster Detection, Optimization, and Interpretation Approach for Financial Data," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13848–13861, Dec. 2022, doi: 10.1109/TCYB.2021.3109066.

[12] G. Kou, Y. Peng, and G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods," *Information Sciences*, vol. 275, no. August, pp. 1–12, Aug. 2014, doi: 10.1016/j.ins.2014.02.137.

[13] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, no. August, pp. 301–315, Aug. 2019, doi: 10.1016/j.eswa.2019.02.033.

[14] G. Wang, F. Li, P. Zhang, Y. Tian, and Y. Shi, "Data Mining for Customer Segmentation in Personal Financial Market," in *Rough Sets and Knowledge Technology*, Lecture Notes in Computer Science, vol. 5589, Berlin, Germany: Springer, 2009, pp. 614–621, doi: 10.1007/978-3-642-02298-2_90.

[15] C. Yang, G. Liu, C. Yan, and C. Jiang, "A clustering-based flexible weighting method in AdaBoost and its application to transaction fraud detection," *Science China Information Sciences*, vol. 64, no. 12, Art. no. 222101, 2021, doi: 10.1007/s11432-019-2739-2.

[16] F. Baser, O. Koc, and A. S. Selcuk-Kestel, "Credit risk evaluation using clustering based fuzzy classification method," *Expert Systems with Applications*, vol. 223, no. August, Art. no. 119882, 2023, doi: 10.1016/j.eswa.2023.119882.

[17] S. Dzuba and D. Krylov, "Cluster analysis of financial strategies of companies," *Mathematics*, vol. 9, no. 24, pp. 1–21, Dec. 2021, doi: 10.3390/math9243192.

[18] S. Becirovic, E. Zunic, and D. Donko, "A Case Study of Cluster-based and Histogram-based Multivariate Anomaly Detection Approach in General Ledgers," in *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*, IEEE, 2020, pp. 1–6, doi: 10.1109/INFOTEH48170.2020.9066333.

[19] G. Herman, B. Zhang, Y. Wang, G. Ye, and F. Chen, "Mutual information-based method for selecting informative feature sets," *Pattern Recognition*, vol. 46, no. 12, pp. 3315–3327, Dec. 2013, doi: 10.1016/j.patcog.2013.04.021.

[20] T. G. Penkova, "Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety," *Procedia Computer Science*, vol. 112, pp. 99–108, 2017, doi: 10.1016/j.procs.2017.08.179.

[21] A. Bahl, B. Hellack, M. Balas, A. Dinischiotu, M. Wiemann, J. Brinkmann, A. Luch, B. Y. Renard, and A. Haase, "Recursive feature elimination in random forest classification supports nanomaterial grouping," *NanoImpact*, vol. 15, Art. no. 100179, pp. 1–12, 2019, doi: 10.1016/j.impact.2019.100179.

[22] C. Barile, C. Casavola, G. Pappalettera, and V. Paramsamy Kannan, "Laplacian score and K-means data clustering for damage characterization of adhesively bonded CFRP composites by means of acoustic emission technique," *Applied Acoustics*, vol. 185, Art. no. 108425, pp. 1–12, 2022, doi: 10.1016/j.apacoust.2021.108425.

[23] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, no. 1, pp. 1157–1182, 2003.

[24] M. Al Fatih Abil Fida, T. Ahmad, and M. Ntahobari, "Variance Threshold as Early Screening to Boruta Feature Selection for Intrusion Detection System," in *2021 13th International Conference on Information & Communication Technology and System (ICTS)*, IEEE, 2021, pp. 46–50, doi: 10.1109/ICTS52701.2021.9608852.

[25] Y. Siti Ambarwati and S. Uyun, "Feature Selection on Magelang Duck Egg Candling Image Using Variance Threshold Method," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, 2020, pp. 694–699, doi: 10.1109/ISRITI51436.2020.9315486.

[26] S. Yu and H. Zhao, "Rough sets and Laplacian score based cost-sensitive feature selection," *PLoS ONE*, vol. 13, no. 6, pp. 1–23, 2018, doi: 10.1371/journal.pone.0197564.

[27] R. Shang, W. Zhang, M. Lu, L. Jiao, and Y. Li, "Feature selection based on non-negative spectral feature learning and adaptive rank constraint," *Knowledge-Based Systems*, vol. 236, no. January, Art. no. 107749, 2022, doi: 10.1016/j.knosys.2021.107749.

[28] D. García-García and R. Santos-Rodríguez, "Spectral Clustering and Feature Selection for Microarray Data," in *2009 International Conference on Machine Learning and Applications*, IEEE, 2009, pp. 425–428, doi: 10.1109/ICMLA.2009.86.

[29] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *Journal of Machine Learning Research*, vol. 5, no. 1, pp. 1205–1224, 2004.

[30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.

[31] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[32] S.-L. Huang, X. Xu, and L. Zheng, "An Information-Theoretic Approach to Unsupervised Feature Selection for High-Dimensional Data," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 157–166, May 2020, doi: 10.1109/JSAIT.2020.2981538.

[33] A. Rizwan, N. Iqbal, A. N. Khan, R. Ahmad, and D. H. Kim, "Toward Effective Pattern Recognition Based on Enhanced Weighted K-Mean Clustering Algorithm for Groundwater Resource Planning in Point Cloud," *IEEE Access*, vol. 9, no. September, pp. 130154–130169, 2021, doi: 10.1109/ACCESS.2021.3111112.

[34] P. Li, C. Wang, J. Wu, and R. Madlenak, "An E-commerce Customer Segmentation Method based on RFM Weighted K-means," in *2022 International Conference on Management Engineering, Software Engineering and Service Sciences (ICMSS)*, IEEE, 2022, pp. 61–68, doi: 10.1109/ICMSS55574.2022.00017.

[35] J.-H. Chen, M.-C. Su, and B. Annuerine Badjie, "Exploring and weighting features for financially distressed construction companies using Swarm Inspired Projection algorithm," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 376–389, Aug. 2016, doi: 10.1016/j.aei.2016.05.003.

[36] A. Kukkar et al., "ProRE: An ACO-based programmer recommendation model to precisely manage software bugs," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 1, pp. 483–498, 2023, doi: 10.1016/j.jksuci.2022.12.017.

[37] Madhumita and S. Paul, "A Feature Weighting-Assisted Approach for Cancer Subtypes Identification From Paired Expression Profiles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 3, pp. 1403–1414, May 2022, doi: 10.1109/TCBB.2020.3041723.

[38] M. Rafi, H. Khan, H. Nadeem, and H. Shakeel, "Unsupervised Topic Aware Document-Level Semantic Representation for Document Clustering," in *2021 22nd International Arab Conference on Information Technology (ACIT)*, IEEE, 2021, pp. 1–10, doi: 10.1109/ACIT53391.2021.9677217.

[39] A. Aradnia, M. A. Haeri, and M. M. Ebadzadeh, "Adaptive Explicit Kernel Minkowski Weighted K-means," *Information Sciences*, vol. 584, no. January, pp. 503–518, 2022, doi: 10.1016/j.ins.2021.10.048.

[40] K. Balaji, "Machine learning algorithm for feature space clustering of mixed data with missing information based on molecule similarity," *Journal of Biomedical Informatics*, vol. 125, no. January, Art. no. 103954, 2022, doi: 10.1016/j.jbi.2021.103954.

[41] A. S. Abdulameer, S. Tiun, N. S. Sani, M. Ayob, and A. Y. Taha, "Enhanced clustering models with wiki-based k-nearest neighbors-based representation for web search result clustering," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 3, pp. 840–850, 2022, doi: 10.1016/j.jksuci.2020.02.003.

[42] A. Benkessirat and N. Benblidia, "A novel feature selection approach based on constrained eigenvalues optimization," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, Part A, pp. 4836–4846, 2022, doi: 10.1016/j.jksuci.2021.06.017.

 ISSN 2723-6471

[43] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, vol. 133, no. September, pp. 285–296, 2021, doi: 10.1016/j.jbusres.2021.04.070.

[44] S. Chowdhury, N. Helian, and R. Cordeiro de Amorim, "Feature weighting in DBSCAN using reverse nearest neighbours," *Pattern Recognition*, vol. 137, no. May, Art. no. 109314, 2023, doi: 10.1016/j.patcog.2023.109314.

[45] Y. Miao and Y. Xu, "A K-Means-Based Interpolation Algorithm With Lp-Norm and Feature Weighting," *IEEE Access*, vol. 12, no. July, pp. 96179–96192, 2024, doi: 10.1109/ACCESS.2024.3424265.

[46] X. Ai and Q. Ji, "Research on Gas User Clustering Algorithm: Based on PCA and Attribute Weighting," in *Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics (ICCIR '22)*, New York, NY, USA: Association for Computing Machinery, vol. 2022, no. October, pp. 783–788, 2022, doi: 10.1145/3548608.3559307.

[47] M. Landauer, F. Skopik, M. Wurzenberger, and A. Rauber, "Dealing with Security Alert Flooding: Using Machine Learning for Domain-independent Alert Aggregation," *ACM Transactions on Privacy and Security*, vol. 25, no. 3, Art. no. 18, pp. 1–36, Apr. 2022, doi: 10.1145/3510581.

[48] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, Feb. 2020, doi: 10.1007/s10462-019-09682-y.

[49] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, Art. no. 1295, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[50] R. C. de Amorim, "Feature Relevance in Ward's Hierarchical Clustering Using the Lp Norm," *Journal of Classification*, vol. 32, no. 1, pp. 46–62, Apr. 2015, doi: 10.1007/s00357-015-9167-1.

[51] T. Li, A. Rezaeipanah, and E. M. Tag El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 6, pp. 3828–3842, Jun. 2022, doi: 10.1016/j.jksuci.2022.04.010.