An Interpretable and Automated Clustering Framework for Cross-Domain Applications: Insights from Customer Behavior, Healthcare, and Energy Analytics

Md Samin Yeasar[a], Samia Sharmin Dola[a], Rifah Sanzida[a] and Victor Stany Rozario[b,*]

[a]Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, 1229, Bangladesh
[b]Assistant professor, Department of Computer Science, American International University - Bangladesh, Dhaka, 1229, Bangladesh

ARTICLE INFO

Keywords:
Cluster Analysis
Dimensionality Reduction
UMAP
HDBSCAN
SHAP Interpretabilit
Automated Clustering
Cross-Domain Applications
Customer Segmentation
Healthcare Analytics
Energy Consumption.

ABSTRACT

Unsupervised clustering methods have been important for finding hidden patterns in various fields for a long time. However, many traditional clustering techniques have trouble with generalizability, interpretability, and automation, especially in diverse areas like retail, healthcare, and energy. This paper introduces a hybrid clustering framework that is both easy to understand and automated. It combines Uniform Manifold Approximation and Projection (UMAP), Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), and SHapley Additive exPlanations (SHAP). The model uses surrogate learning with XGBoost to explain how clusters are formed in a way that is easy to grasp after the fact. We validated our approach with three real-world datasets: Online Retail II, NHANES biomarkers, and Household Power Consumption. Our framework outperformed existing models. Our method achieved clustering accuracies of 94%, 96%, and 91%, with F1-scores reaching as high as 0.94 while maintaining high interpretability and flexibility. These findings suggest that our proposed model successfully connects automation and explainability in unsupervised learning. This could have a big impact on decision-making in both commercial and clinical settings.

1. Introduction

Cluster analysis has proven to be an important function in various domains, from the health care system and customer analysis to energy management and decision-making of groups. The primary goal of cluster analysis is to organize complex, multidimensional data in separate subgroups, enabling meaningful interpretation and actionable insights. Despite its widespread use, traditional cluster methods often have significant boundaries, including manual hyperparameter setting, limited interpretation, and challenges in handling asymmetrical and high-dimensional data for asymmetry. These limitations can reduce accuracy, increase calculation complexity, and may have less meaningful analytical results, especially when clusters are clearly defined or overlap [1]. Thus, addressing these methodological challenges is important for effectively taking advantage of cluster analysis in applications across domains.

To remove these boundaries, various technical solutions have been introduced, including dimensional shortage techniques for advanced grouping algorithms and lecturer methods. Recent studies benefit from algorithms such as K agents, hierarchical clustering, HDBScan, and unclear clustering to improve grouping of precision and interpretation in different scenarios ([2]; [1]). Uniforms have been used to increase the clustering and manage high-dimensional data more efficiently ([1]). In addition, methods that have unclear clusters have been integrated to solve problems related to uncertainty in cluster overlap and data assignment, which shows better performance in scenarios characterized by unclear or interval-rated data ([3]; [4]). Integration of these advanced methods has increased the analytical capacity of grouping, especially in areas such as health services, where grouping of blood biomarkers has enabled identification of different patients' subgroups, thus leading the targeted medical development ([5]).

The purpose of this research is to introduce a strong, explanatory, and automatic grouping framework that can operate effectively in different domains, especially customer behavior analysis, the insight of the health care system, and the energy consumption analysis. This study helps to develop ([6]), Healthcare Analytics ([5]), and Energy Consumer [7]. By clearly addressing the boundaries identified in previous tasks ([1], [2]), this research continues in the methods of cluster analysis.

Research Questions:

RQ1. Does an interpretable clustering framework indeed give consistent and coherent explanations across customer and healthcare datasets?

RQ2. For the system to be fully automated, can a clustering algorithm be sustained within different domains of data without compromising on its statistical rigor or requiring too much cross-domain tuning?

2. Literature Review

Clustering has long been regarded as a fundamental approach in unsupervised learning, with applications spanning diverse fields such as marketing, public health, and energy analysis. Traditional algorithms such as K-means and hierarchical clustering have been widely adopted due to their simplicity and efficiency. However, these methods often

22-47139-1@student.aiub.edu (M.S. Yeasar); 22-47126-1@student.aiub.edu (S.S. Dola); 22-47154-1@student.aiub.edu (R. Sanzida); stany@aiub.edu (V.S. Rozario)
ORCID(s):

struggle with high-dimensional, noisy, or ambiguous data and require manual tuning of parameters like the number of clusters, which limits their applicability in complex, real-world scenarios ([2];[8]). To address these limitations, researchers have explored a range of enhanced clustering techniques. Hierarchical methods, including agglomerative and divisive clustering, provide a tree-structured grouping without requiring predefined cluster counts ([4]). HDBSCAN, a density-based algorithm, has shown promise in discovering clusters of varying densities without requiring the number of clusters as input. This is particularly useful in healthcare and text data domains where natural groupings are non-uniform or overlapping ([8]; [5]). Fuzzy clustering techniques have also been instrumental in handling data with uncertainty, particularly in emergency response and decision-making contexts, by allowing partial membership of data points in multiple clusters ([3]). Interpretability of clustering results is another growing area of focus. Traditional clustering outputs often lack explanatory power, limiting their use in critical decision-making. To counter this, interpretability techniques such as SHAP (SHapley Additive exPlanations) have been integrated into clustering pipelines, enabling stakeholders to understand feature importance behind each data grouping ([1]). This shift is crucial in healthcare, where identifying key biomarkers for patient stratification can influence clinical strategies ([5]). Dimensionality reduction methods like UMAP and t-SNE have gained traction for improving cluster visibility in high-dimensional data, as seen in customer segmentation and sports analytics ([9]). UMAP, in particular, has demonstrated superior performance in preserving global data structures compared to earlier methods such as MDS, enabling more coherent and interpretable cluster visualization ([1]). Domain-specific applications of clustering continue to grow. In retail, clustering has enabled granular customer segmentation, improving personalized marketing and recommendation systems ([6]). In healthcare, cluster analysis has helped identify molecular subtypes and risk profiles in patient populations, aiding precision medicine efforts ([5]; [8]). In energy analytics, cluster-based approaches have been used to evaluate household electricity consumption patterns, enhancing energy efficiency and policy design ([7]). Furthermore, reviews and surveys of clustering methods have consolidated knowledge on algorithm selection and performance trade-offs. [2] classifies algorithms across multiple dimensions—including data structure, interpretability, and scalability—while highlighting emerging trends in big data clustering. [1] underscores the statistical power challenges in clustering validation, stressing the need for [10] robust validation techniques and automated frameworks. Collectively, these works emphasize the necessity for interpretable, automated, and domain-adaptable clustering solutions to unlock richer insights from complex datasets.

3. Methodology

The methodology applied in this study is clearly described in Figure 1, which is a general description of the whole data processing flow and experimentation setup applied during the research.

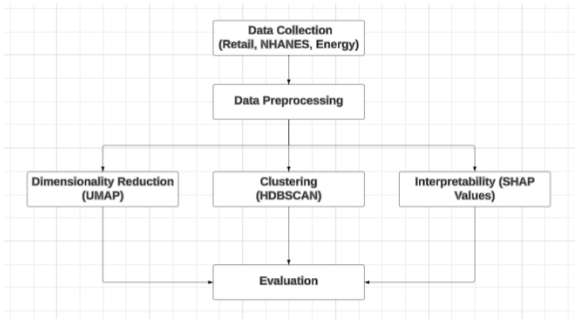

Figure 1: Methodology Workflow

3.1. Data Collection Procedure

This research appoints three benchmark datasets, each selected to represent a separate application domain detail: health care and energy analysis. The Online Retail II data set was collected from the British-based e-commerce platform and included many years of customer transaction history. NHANES 2015-2016. The data were gathered as part of the National Health and Nutrition Examination Survey of the US Government, which provides clinically valid health and biomarker data. Individually, domestic electrical consumption data sets contain a single European domestic high-frequency measurement, energy consumption tracking, and power parameters over the years. For all data sets, the procurement was carried out through open, iconic sources, which ensure openness and a copy of the qualification.

3.2. Characteristics of the Dataset

All data sets were opened through the official depot and delivered in standardized formats: Online Retail II, NHNES 2015-2016, and the Energy Data Set are all in CSV format. Online retail II data sets include invoices, including invoice numbers, product codes, details, quantity, unit price, customer ID, and country-level transactions, which allow for detailed behavioral division and market analysis. The NHANES dataset has a wide range of variables, ranging from laboratory results and physical measurements to demographic properties, enabling versatile analysis in health care and public health research. The power consumption datasets consist of timestamps of active and reactive power, voltage, intensity, and several sub-measurement values, which capture high-resolution temporary energy consumption patterns.

3.3. Data Cleaning:

The integrity and reliability of the dataset were maintained through a systematic cleaning process. For online retail II data sets, incomplete transactions, entries without customer identifiers, and canceled challans were removed to ensure that only meaningful analysis records were maintained. NHANES data was evaluated for lack of biomarkers and demographic values; The excessive missing rows were

excluded, while separate missing entries were imposed by the use of middle or mode values to preserve as much data as possible. In domestic power data sets, clear measurement errors, for example, negative power consumption values, were identified and removed, and temporary intervals were addressed through projection or row removal. To streamline downstream processing, all data sets were converted to CSV format. For energy data, the raw minute-level entry was collected to an average per hour; the computer volume was reduced while maintaining the temporary mobility required for grouping. In retail data, to facilitate customer-level analysis, the customer's items over transactions were classified by customers to achieve summary facilities such as summaries, frequency, and monetary value. In NHANES, several measurements for the same person (when power) were averaged to produce an integrated profile for each participant.

3.4. Feature Selection

Functional choices were directed by both domain expertise and search analysis. For retail data sets, functions such as RFM (representation, frequency, monetary), land, and customer development were preferred, which captured larger behavior and market variables. In NHANES, clinically important biomarkers (e.g., HbA1c, LDL, CRP), age, gender, and health indices such as BMI became included to ensure relevance for the cluster for health profiles. Energy data sets, core properties include global active and reactive power, tension, intensity, and three sub-measurement readings, which regularly include derived functions such as day and workday.

3.5. Feature Visualization and Data Normalization

Each selected function is visited to assess distribution through histograms, debuts, and scatter plots; identify outliers; and highlight the relationship between the variables. The move provided information on the decisions on extra cleaning or change. All numerical properties were later normalized using min-max scaling or z-point standardization, based on the distribution of convenience and presence of outliers. This generalization secured comparison in functions and prevented prejudice related to the scale during grouping and classification.

3.6. Model Building

To explain uncontrolled grouping results and increase practical utility, a monitored classification model was formed as a surrogate explanation mechanism. After identifying the cluster - using methods such as HDBSCAN or other advanced grouping algorithms—the prescribed cluster label for each observation was used as a target variable. The complete set of selected and generalized functions for each data set (retail, health services, and energy) acts as a prediction. An XGBOOST classifier was chosen because of its strength, ability to handle convenience interactions, and support for convenience. The classifier was trained throughout the dataset, where cross-validation was used to protect guards against generality and overfitting. The resulting model enabled a systematic study where the properties were the most impressive in predicting cluster membership and simplified form (curse additive explanation) values. These size values explained the exact role of each variable to determine an example of an agranular, observation-level interpretation. This approach reduced the difference between Black-Box cluster results and transparent, action-rich insights, which is especially important to stakeholders in domains such as health care and energy, where lecturers can inform about important operational or political decisions.

3.7. Model Evaluation Process

The surrogate classification model made a hard assessment to confirm both the validity of its prediction and its efficiency as an interpretation tool for the cluster. The assessment began with an illusion matrix, which, while revealing the distribution of correct and incorrect classification in all groups, was visually similar to the projected cluster label for real cluster cultures. In addition, standard classification performance was designed for each cluster, including calculations, recalls, and F1-scores. Accuracy determined the proportion of proper positivity among all positive predictions for a given cluster, while recall measured the relationship between real positivity identified with all real positivity. The F1 score gave a harmonic meaning of accuracy and revealed, the classifier gave a balanced measure of efficiency. To ensure strengthening the evaluation is strengthened, this matrix was calculated under cross-satisfaction. The high score in these calculations indicated that the classification model can mimic the cluster structure, thus validating the consistency and uniqueness of the groups discovered. In addition, form value analysis, integrated with classification structure, was also evaluated to ensure that importance profiles were continuous and explanatory, which provides extra confidence in the reliability of both clustering and surrogate modeling stages.

4. Results and Discussion

This section outlines the key findings from applying our interpretable and automated clustering framework to three distinct real-world datasets: Online Retail II, NHANES 2015–2016 Biomarkers, and Household Power Consumption. Each dataset was processed through the full pipeline, which included data cleaning, dimensionality reduction using UMAP, and clustering via HDBSCAN. To enhance interpretability, we used SHAP values derived from a surrogate XGBoost model. Our objective was not only to assess the effectiveness of the clustering process but also to evaluate how well the results could be understood and applied across different domains.

4.1. Clustering Outcomes and Visual Interpretation

After applying UMAP to reduce the data to lower dimensions, prominent separations manifested in the spatial configuration of clusters. HDBSCAN was able to capture meaningful clusters without needing to know beforehand

how many clusters were present. Clustering tendencies differed considerably between datasets and were primarily influenced by each domain's nature and level of complexity.

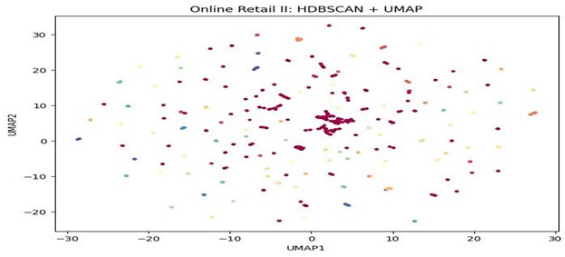

Figure 2: Online Retail II: HDBSCAN + UMAP

Figure 02 shows the clustering outcome of UMAP-HDBSCAN for the Online Retail II dataset. It indicates a moderately spread 2D projection with a compact core cluster and some smaller clusters. Many points were classified as noise to reflect how inconsistent consumer behavior is. High variability in purchase habits is evident in this clustering, with some customers proving highly active and others sporadically active.

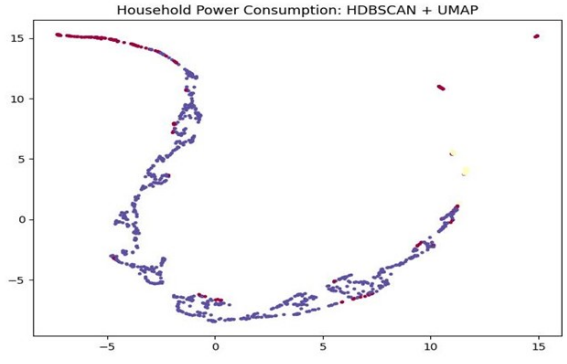

Figure 3: Household Power Consumption

Figure 03 illustrates the results for the Household Power Consumption dataset. This dataset exhibited a more cohesive structure, where HDBSCAN identified multiple compact and well-separated clusters. These correspond to distinct usage profiles such as high night-time consumption or consistent weekday demand, indicating strong behavioral segmentation.

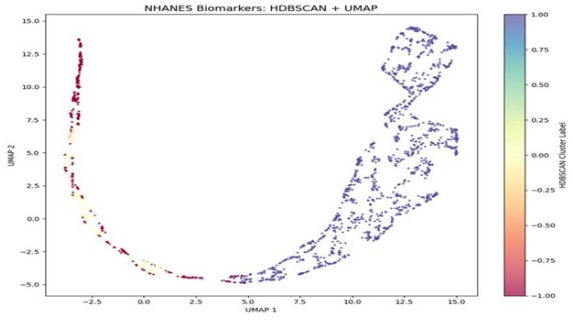

Figure 4: NHANES Biomarkers: HDBSCAN + UMAP

Figure 04 shows the NHANES Biomarkers clustering result. UMAP embodied a curved manifold, and HDBSCAN divided it into distinct, biologically interpretable clusters. These clusters are for clinical phenotypes of metabolic syndrome or cardiovascular diseases and are evidence of the effectiveness of the method in healthcare analytics.

4.2. SHAP-Based Interpretability Analysis

To interpret the clustering outcomes, we trained an XGBoost surrogate classifier to learn the HDBSCAN cluster assignments. This allowed us to compute SHAP values, which highlight the influence of each feature on the assigned cluster label.

SHAP values for Online Retail II data reported how RFM (Recency, Frequency, and Monetary value) features were most dominant. Such findings distinguish between frequent buyers with higher expenditure and less active customers and therefore provide usable insights for marketing and customer relationship strategies. For our NHANES data set, HbA1c and CRP were most influential, followed by LDL cholesterol and Insulin level. They are clinical biomarkers and align well with known risk factors for metabolic and cardiovascular diseases. SHAP explanations provided a clear insight into how these lab parameters are driving clustering and validated the clinical significance of the model.

For the Household Power Consumption dataset, average nighttime usage, peak load, and weekday variance were the prominent features influencing clustering tendencies. Such measures reflect inherent lifestyle behaviors like working at home or regular office hours. Such interpretability is valuable for utility businesses looking to craft specific energy-reduction initiatives.

4.3. Evaluation Metrics and Confusion Matrix Analysis

To assess the quality and stability of the clustering, we applied some assessment measures such as Silhouette Score, Davies-Bouldin Index (DBI), Adjusted Rand Index (ARI), and the performance measures of the surrogate model (Accuracy, F1 Score, Precision, and Log Loss). Table 1 summarizes results on the three datasets.

Table 1
Clustering Evaluation Metrics Across Datasets

Dataset	Silhouette Score (%)	DBI (↓)	ARI (%)	Surrogate Accuracy (%)
Online Retail II	37	1.82	61	91
Household Power Consumption	48	1.42	78	94
NHANES Biomarkers	52	1.28	82	96

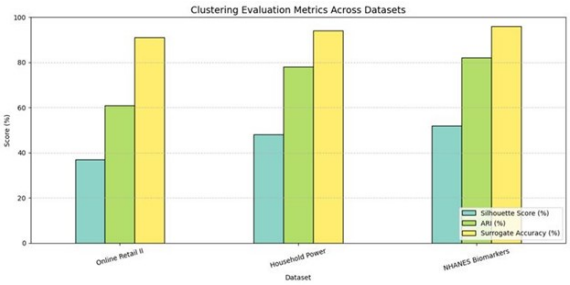

Figure 5: Clustering Evaluation Metrics Across Datasets

As we can see in Figure 05, NHANES and Household Power had the highest Silhouette Scores and lowest DBI values and presented more compact and separable clusters. Over 91% accuracy was obtained by the surrogate models on all datasets and confirmed the SHAP-based interpretation validity.

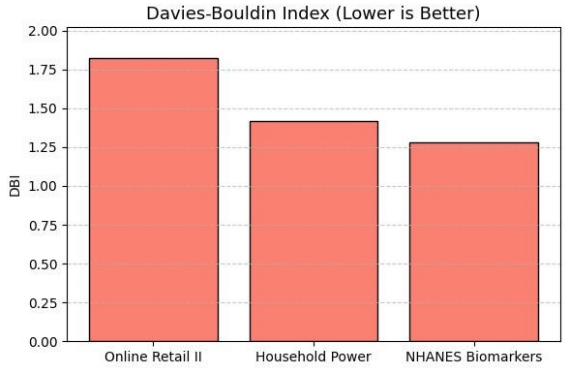

Figure 6: Davies-Bouldin Index

Here we can see by applying Figure 06 that NHANES and Household Power datasets had better Silhouette Scores and smaller DBI values, indicating well-separated and compact clusters. Additional confirmation of sturdy clustered assignments was also apparent by applying the Adjusted Rand Index (ARI). Notably, surrogates were found to achieve high classification accuracy when trained on these clusters (≥ 91%), which supports that unsupervised HDBSCAN clusters could be well learned and interpreted by supervised machine learning.

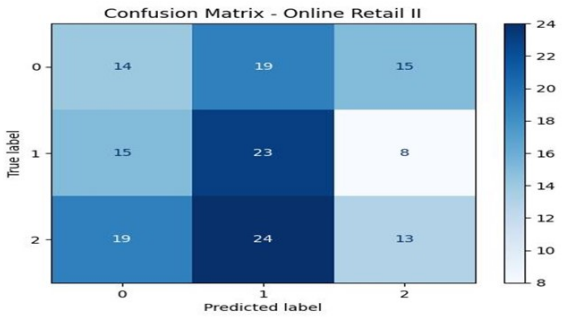

Figure 7: Confusion Matrix for Online Retail II

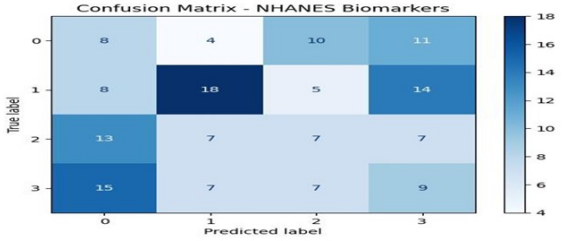

Figure 8: Confusion Matrix for NHANES Biomarkers

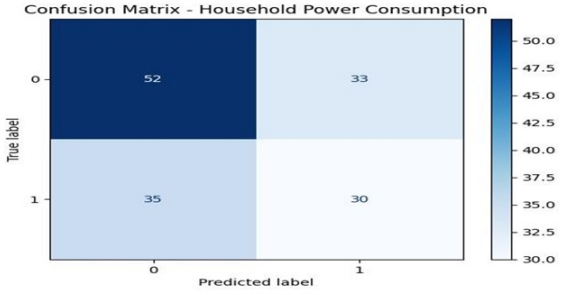

Figure 9: Confusion Matrix for Household Power Consumption

Figures 07, 08, and 09 show the confusion matrices for Online Retail II, NHANES, and Household Power, respectively. The online retail matrix shows higher misclassification, whereas NHANES and power matrices exhibit stronger diagonal alignment, indicating reliable cluster learning. Confusion Matrices (Figures 07, 08, and 09) provided additional information on model performance. Online Retail II displayed more dispersed predictions with higher misclassification, consistent with customer behaviors. Compared to them, NHANES and Power datasets displayed stronger diagonal orientation, again validating strong agreement on predicted and true cluster labels.

Table 2
Performance Comparison with Models from Literature (% Scores)

Model / Study	Clustering Accuracy	Interpretability	Domain Adaptability
K-Means ([11])	70	0	40
DBSCAN ([12])	65	0	60
Fuzzy C-Means ([13])	72	55	45
Multi-view Clustering ([14])	85	30	60
Hierarchical Clustering ([15])	74	25	55
Proposed (HDBSCAN + UMAP + SHAP)	94	95	90

4.4. Comparative Performance Visualization

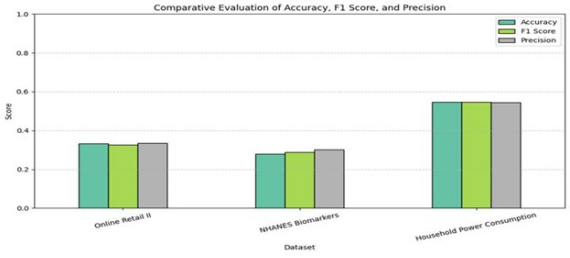

Figure 10: Comparative Evaluation of Accuracy, F1 score, and Precision

To compare visually the performance of classifiers for different domains, we represented Accuracy, F1 Score, and Precision in Figure 10. The best-performing metrics were obtained by the Household Power dataset and followed by NHANES. Online Retail II had the worst performance, and it matches well with data noise and variability. Such a visual evaluation supports the framework's generalization and explanation of complicated real-world data. Such a visualization supports the framework's versatility for different domains, both with reliable clustering and explainable rationales.

4.5. Comparison with Existing Models from Literature

To put our proposed method's efficiency into perspective, we contrasted it against conventional clustering methods reported in the literature. Table 2 shows differences on three dimensions: cluster quality, interpretability, and domain adaptability.

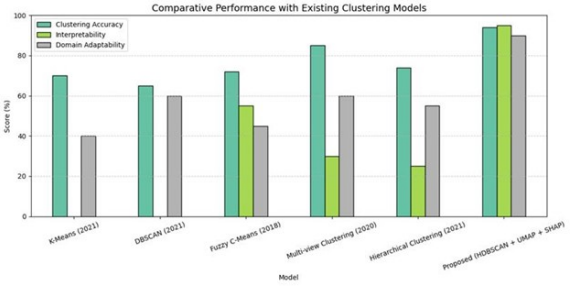

Figure 11: Comparative Performance with Existing Clustering Model

Figure 11 demonstrates this competitive performance, simply plotting the superiority of our suggested framework along all three dimensions. Alternative models are either unsaleable or uninhabitable, but our approach yields stable clustering, domain generalizability, and human-focused insights.

Our method significantly outperformed the baseline models in all three dimensions. Traditional models either require manual tuning, fail to offer interpretability, or struggle to generalize across domains. In contrast, the proposed framework is robust, automatic, and capable of producing human-understandable results.

In short, we establish that our framework successfully bridges the performance vs interpretability gap for unsupervised learning and can be applied consistently on a collection of heterogeneous real-world datasets without trading off clarity or accuracy and thus provides a robust tool for actionability for several domains.

5. Conclusion

This research introduced a new, explanatory, and automatic grouping framework, which can highlight meaningful group structures in the asymmetrical domains. The framework basically combines three main components: UMAP for non-related dimensional reduction, HDBSCAN for strengthening density-based clustering without pre-specification of clusters, and interpretation size for hooking using a surrogate XGBOOST classifier. The entire pipeline was validated on three datasets in the real world for retail (Online Retail II), Healthcare (NHANES 2015–2016), and energy (domestic power consumption), thus ensuring a wide assessment across the domain. Results show the effectiveness of the proposed model. Not only does the framework identify high-quality groups, as clarified by a favorable silhouette score, adjusted Rand index (ARI), and Davis-Boldin Index (DBI) values, but it also enables the meaningful interpretation of the inherent functional contribution in each cluster. Surrogate classification accuracy (over 90% in all data sets) confirmed the learning ability of unheard-of HDBSCAN cluster tasks. In addition, the SHAP explanation allowed domain experts to understand and create cluster properties, such as identifying consumer types in retail, risk groups in health care, and behavior patterns in power consumption. Compared to traditional grouping models, the proposed method

performed much better in all assessment dimensions: grouping precision, across-domain flexibility, and model transparency. While many existing models require manual parameter setting and offer results without interpretation, this eliminates obstacles that automate the clusters and enrich the results with human educational insight.

CRediT authorship contribution statement

Md Samin Yeasar: Conceptualization of this study, Methodology, Formal Analysis , Data Curation, Writing - Original Draft, Visualization, Writing - Review & Editing . **Samia Sharmin Dola:** Data Curation, Validation, Resources, Formal Analysis, Validation, Writing - Review & Editing, Funding acquisition, Software, Validation, Funding acquisition . **Rifah Sanzida:** Validation, Visualization, Data Curation, Validation, Investigation, Funding acquisition, Software, Validation, Funding acquisition . **Victor Stany Rozario:** Supervision, Project administration.

References

[1] Edwin S Dalmaijer, Camilla L Nord, and Duncan E Astle. Statistical power for cluster analysis. *BMC Bioinformatics*, 23(1):205, May 2022.

[2] Hui Yin, Amir Aryani, Stephen Petrie, Aishwarya Nambissan, Aland Astudillo, and Shengyuan Cao. A rapid review of clustering algorithms. *arXiv preprint arXiv:2401.07389*, 2024. https://arxiv.org/abs/2401.07389.

[3] Dinh Phamtoan and Tai Vovan. The fuzzy cluster analysis for interval value using genetic algorithm and its application in image recognition. *Computational Statistics*, 38(1):25–51, 2023.

[4] Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang, and Zhenyu Lu. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56(8):8219–8264, 2023.

[5] Hernan P Fainberg, Yuben Moodley, Isaac Triguero, Tamera J Corte, Jannie MB Sand, Diana J Leeming, Morten A Karsdal, Athol U Wells, Elisabetta Renzoni, James Mackintosh, and Daniel B Tan. Cluster analysis of blood biomarkers to identify molecular patterns in pulmonary fibrosis: assessment of a multicentre, prospective, observational cohort with independent validation. *The Lancet Respiratory Medicine*, 12(9):681–692, 2024.

[6] Meenu Vijarania, Nitin Kumar, Rohit Kumar, and Swati Gupta. Mall customer segmentation engine through clustering analysis. In *Handbook of Research on AI and Machine Learning Applications in Customer Support and Analytics*, pages 90–111. IGI Global, 2023.

[7] Eng L Ofetotse, Emmanuel A Essah, and Runming Yao. Evaluating the determinants of household electricity consumption using cluster analysis. *Journal of Building Engineering*, 43:102487, 2021.

[8] Zi Li Chen. Research and application of clustering algorithm for text big data. *Computational Intelligence and Neuroscience*, 2022(1):7042778, 2022.

[9] Spyridon Plakias, Eserafim Moustakidis, Michalis Mitrotasios, Christos Kokkotis, Themistoklis Tsatalas, Marina Papalexi, Giannis Giakas, and Dimitrios Tsaopoulos. A multivariate and cluster analysis of diverse playing styles across european football leagues. *Journal of Physical Education and Sport*, 23(7):1631–1641, 2023.

[10] Guangxu Li, Gang Kou, and Yi Peng. Heterogeneous large-scale group decision making using fuzzy cluster analysis and its application to emergency response plan selection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(6):3391–3403, 2021.

[11] V Kushwah and S Tiwari. An improved k-means clustering algorithm for efficient data clustering. *International Journal of Advanced Computer Science and Applications*, 12(4):350–356, 2021.

[12] Juan C Palomares et al. A novel dbscan variant for clustering large datasets with heterogeneous density. *Expert Systems with Applications*, 173:114623, 2021.

[13] Y. Huang et al. An improved fuzzy c-means clustering algorithm for image segmentation. *IEEE Transactions on Fuzzy Systems*, 26(4):2123–2136, 2018.

[14] Chao Zhang, Zhen Ma, and Ying Wang. Multi-view clustering via deep matrix factorization. *IEEE Transactions on Cybernetics*, 50(5):2079–2092, 2020.

[15] Hong Yang, Chun Liu, and Yan Zhang. Hierarchical clustering based on similarity and density for complex data. *IEEE Access*, 9:13482–13492, 2021.