Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 1, 19–31
DOI: 10.61822/amcs-2025-0002

ON EXPLAINABILITY OF CLUSTER PROTOTYPES WITH ROUGH SETS: A CASE STUDY IN THE FMCG MARKET

MAREK GRZEGOROWSKI [a,*], ANDRZEJ JANUSZ [a,b], ŁUKASZ MARCINOWSKI [c], ANDRZEJ SKOWRON [d], DOMINIK ŚLĘZAK [a,e], GRZEGORZ ŚLIWA [c]

[a] Institute of Informatics
University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
e-mail: {m.grzegorowski, janusza, slezak}@mimuw.edu.pl

[b] School of Information Systems
Queensland University of Technology
Gardens Point Campus, Brisbane, Australia

[c] FitFood
Solskiego 11/28, 31-216 Cracow, Poland
e-mail: {g.sliwa, l.marcinowski}@fitfoodpoland.pl

[d] Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland

[e] QED Software
ul. Mazowiecka 11/49, 00-052 Warsaw, Poland
e-mail: dominik.slezak@qedsoftware.com

Despite the growing popularity of machine learning (ML), such solutions are often incomprehensible to employees and difficult to control. Addressing this issue, we discuss some essential problems of explainable ML applications in the fast-moving consumer goods (FMCG) market. This research puts forward a new approach to effective supply management by utilizing rough sets (RST), distance-based clustering, and dimensionality reduction techniques. In the presented case study, we aim to reduce the work done by experts by applying a single delivery plan to many similar points of sale (PoS). We achieve this objective by clustering vending machines based on historical sales patterns. To verify the feasibility of such an approach, we performed a series of experiments related to demand prediction on two data representations with various clustering techniques. The conducted experiments confirmed that, without losing quality in terms of MAE and RMSE, we could operate on PoS in an aggregate manner, thus reducing the workload of preparing delivery plans.

Keywords: RST, clustering, PCA, UMAP, XAI, LLM, TRISM, FMCG, supply management.

1. Introduction

The recent progress in machine learning (ML) triggered rapid technological advances in many sectors. One of the beneficiaries of this technological transformation is the FMCG industry. For instance, for food producers and distributors, the demand misestimation may lead to lost sales opportunities or food waste, thus the successful application of ML models to demand estimation is a competitive factor for many manufacturers and retailers (Tarallo *et al.*, 2019; Malefors *et al.*, 2021). A special case, requiring even more attention, is the distribution of meals via vending machines, where the applications of artificial intelligence (AI) techniques may lead to the optimization of operational processes related to supply

[*] Corresponding author

management (Grzegorowski *et al.*, 2022).

However, despite the growing popularity of AI/ML, the adoption of intelligent systems is very limited and faces many impediments. Such solutions are often incomprehensible to employees and difficult to control, causing reluctance among staff. The limited trust and overall complexity of ML algorithms and deep models are the primary factors impeding the widespread integration of ML technologies, holding back their popularization (Adadi and Berrada, 2018). The above-mentioned concerns are just a sample of many challenges related to tackling trust, risk, and security in AI models (TRISM) and achieving more interpretable and explainable AI (XAI) (Habbal *et al.*, 2024; Dwivedi *et al.*, 2023).

To overcome the above-mentioned limitations, we propose a novel approach to apply granular computing (GrC) and unsupervised ML methods in an understandable, user-centric way. In particular, we apply clustering to build granules of similar objects and identify a prototype of each, i.e., the most representative element. In the proposed framework, we interact with experts, who perform their activities (in the discussed case study related to supply management) but only for the selected prototypes, which significantly limits their effort since for the rest of similar elements we can propagate the expert-based solution automatically.

To make the task more comprehensible, we apply rough set reducts (Pawlak, 1982) to identify the relation between clusters and their most distinguishable characteristics to construct human-readable clusters' descriptions using templates of sentences filled with the identified characteristics. The additional component of the process allows rephrasing the descriptions with the preferred large language model (LLM). The developed process provides several measures related to trustworthiness, for example, monitoring of demand estimation with standard regression measures like mean average error (MAE). Moreover, we monitor the stability of the clustering by providing a two-dimensional visualization of week-to-week deviations with uniform manifold approximation and projection (UMAP) and principal component analysis (PCA).

The presented case study concerns the application of ML to optimize supply management and delivery planning for `FitBoxY.com` that distributes ready-to-eat lunch meals with short expiry dates through a large network of developed smart vending machines. To avoid food expiration, the demand at each point of sale has to be carefully estimated, considering many factors, like weather forecasts, calendar events, and historical customers' purchase patterns. The application of cloud computing and ML allowed `FitBoxY.com` to develop a fully automatic supply management system with prescriptive analytics capabilities, as described by Grzegorowski *et al.* (2022), where XGBoost and deep models are applied for demand prediction and heuristic search is applied to generate supply proposals.

However, in the case of frequent rotation of product lines or brands in selected locations, the quality of ML models decreases significantly due to the excessive variety and variability in time series related to sales transactions. This enforces a manual supply process, which has a strong impact on the expected workload. In this study, we evaluate the possibility of optimizing supply management in such situations.

The challenge here is to properly aggregate data in a way that minimizes the variance in purchasing patterns by grouping PoS with similar sales patterns. We examine two different data representations–based on aggregating sales for individual products and their category. We ensure that the results of the data clustering are understandable for the team of experts by referring to XAI prototypes (Heide *et al.*, 2021). Furthermore, we propose an innovative method of generating human-readable cluster descriptions inspired by feature ranking (Adadi and Berrada, 2018), which relies on reduction algorithms from rough set theory (RST) (Pawlak and Skowron, 2007), thereby obtaining a good understanding of each cluster's most discernible characteristics. The conducted research extends our former work (Grzegorowski *et al.*, 2023; 2022) and confirms that without a significant loss of the quality of demand prediction, we can operate on points of sale in an aggregate manner, and thus reduce the amount of work needed to prepare delivery plans. The main contributions of the paper are as follows:

(i) optimization of supply management with GrC,

(ii) clustering stability assessment by the 2D projections with PCA and UMAP,

(iii) a novel approach to interpretability based on RST,

(iv) experimental evaluation of two representations of data from `FitBoXy.com`.

The rest of the paper is organized as follows. In Section 2, we review the literature. Section 3 provides the essential preliminary knowledge about RST, data clustering, and dimensionality reduction. In Section 4, we present the proposed solution. Section 5 describes experimental evaluation. Section 6 summarizes the study.

2. Related works

Food production is a complex process under high uncertainty resulting in differences between planned and actual demand. Considering the short shelf-life of products that may result in food waste, the accurate prediction of the future demand at each point of sale is very important (Tarallo *et al.*, 2019). It is particularly interesting to prepare such a delivery plan for each vending machine, the realization of which will bring maximum profit and minimize food waste. One of the ways is to predict demand with ML models.

State-of-the-art methods such as XGBoost or deep neural networks are very effective in estimating demand (Grzegorowski *et al.*, 2022), yet suffer from a lack of interpretability. On the other hand, models that are intrinsically interpretable, yet simpler, often fall short in accuracy compared with their more sophisticated counterparts (Adadi and Berrada, 2018). This issue becomes particularly problematic in the context of time series data collected from vending machines, which are characteristically very short. Hence, machine learning models need to effectively alleviate the cold-start problem (Kannout *et al.*, 2024).

Maintaining trustworthy human-computer collaboration is a vital research topic. Among the plethora of ML explainability-related methods (Barredo Arrieta *et al.*, 2020), a particularly interesting are model agnostic approaches (Dwivedi *et al.*, 2023). The example-based explanations are explicitly inspired by the cognitive science of human reasoning, which is often prototype-based. For explaining text clusters, keyword extraction seems to be a feasible approach (Penta and Pal, 2021), but this method is not applicable in the general case. Other methods capable of explaining the clusters' similarities are based on variable rankings (Fisher *et al.*, 2019; Zhang *et al.*, 2017). Recent advances in natural language processing and large pre-trained language models are also helpful in improving interactions with users (Min *et al.*, 2023).

In the context of PoS clustering, we require finding a set of important differences between objects from different clusters. Therefore, the application of RST-based reduction methods (Grzegorowski and Ślęzak, 2019; Janusz and Ślęzak, 2015) to facilitate this process, we find a promising approach. Furthermore, considering the variability of sales patterns in time impacting the cluster structure, it is also worth paying attention to the stability of clustering and explanations and various approaches to visual explanation techniques (Barredo Arrieta *et al.*, 2020), particularly 2D projections.

3. Preliminary knowledge

3.1. Rough set theory. Proposed by Pawlak (1982), rough set theory (RST) provides a formalism for reasoning about imperfect data. In RST, objects $u \in U$, where U is a finite, nonempty set, are characterized by their attributes. A finite, non-empty set of attributes is denoted by A. A decision attribute d defines the partitioning of U into disjoint sets corresponding to decision classes (Pawlak and Skowron, 2007). A tuple $(U, A \cup \{d\})$, where $A \cap \{d\} = \emptyset$, is called a decision table and is denoted by $\mathbb{S}$. One may consider $a \in A$, as functions $a : U \to V_a$, where V_a is the set of values of a. Here, a typical way to represent $\mathbb{S}$ is a table with rows corresponding to objects, columns to attributes, and cells to pairs (u, a) assigned values $a(u) \in V_a$.

In RST, a large emphasis is put on the granulation of the attribute space and multivariate feature selection (FS) (Grzegorowski, 2023). A fundamental concept related to FS is a decision reduct $R \subseteq A$, which is an irreducible subset of attributes (features, columns) that determines a decision class (d) at the same level as the whole set of attributes A. In the literature we may find numerous definitions and algorithms allowing calculation of reducts and their approximations (Grzegorowski, 2023; Pięta and Szmuc, 2021; Janusz and Ślęzak, 2015; Stawicki *et al.*, 2017). Approximate reducts are usually based on functions evaluating degrees of information induced by reduced attribute subsets, and may lead to slightly less accurate results, yet could be preferred in some real-life applications when handling huge volumes of data to achieve smaller representations (Grzegorowski and Ślęzak, 2019). For instance, dynamically adjusted approximate reducts (DAAR) (Janusz and Ślęzak, 2015), is a combination of iterative filter-based FS and statistical significance tests. This concept is applied in this study to determine the most distinguishing attributes (cf. Section 4.2) as a special implementation of the variable importance XAI method (Fisher *et al.*, 2019). Our method does not just rank attributes, yet provides a complete subset of descriptive ones.

3.2. Clustering methods. Clustering constitutes an unsupervised learning technique that allows grouping similar objects into the so-called clusters. In the conducted research, we considered several distance-based clustering methods including *kmeans* and partitioning around medoids *pam*. The k initial centers of each cluster in *kmeans* are randomly initiated and iteratively refined in each iteration of the algorithm. Cluster centers may not coincide with an actual data instance; therefore the prototype object of a cluster is chosen as the instance closest to the center. Partitioning around medoids works similarly. Though, in each iteration, a new cluster center (medoid) is selected as the instance with a minimal distance to all other elements. Hence, medoids can be used as the most representative instances (Ikotun *et al.*, 2023).

Agglomerative clustering methods create a hierarchy by starting from singleton groupings, and iteratively merging the two closest groups into a bigger cluster (Kannout *et al.*, 2024). This bottom-up approach ends with all objects consolidated into a single cluster. To measure dissimilarity between groups, agglomerative algorithms utilize linkage functions. In experiments, we evaluated *single_linkage*, *complete_linkage*, and *ward_linkage*. The first one (*single_linkage*) measures the (dis)similarity between two clusters as the smallest distance between any two elements of those clusters; *complete_linkage* measures the largest distance between any two objects; Ward's minimum variance method (*ward_linkage*) relies on the total squared distances between all pairs of instances from clusters. Divisive methods, e.g., divisive analysis (DIANA, *diana_linkage*), construct the hierarchy in the inverse order, starting with objects grouped together, and recursively dividing a group with the largest diameter into two groups whose diameters are possibly small (Ezugwu *et al.*, 2022).

3.3. Dimensionality reduction. Dimensionality reduction techniques are useful for 2D visualizations of high-dimensional data (Zong *et al.*, 2020). There are many dimensionality reduction methods, yet the most prominent in the context of 2D visualization are uniform manifold approximation and projection (UMAP) and principal component analysis (PCA). UMAP is a nonlinear technique, which provides a low-dimensional graph that maintains the relationships existing in the original high-dimensional data in a way that similar objects are typically grouped together. The process consists of two steps. The first involves learning the structure of the manifold, whereas the second prioritizes identifying a low-dimensional representation, by creating a neighbor graph that calculates a similarity score for each point and its neighbors and employs fuzzy cross-entropy to retain similarities of points in the low-dimensional embedding space.

PCA is used to reduce the dimensionality of datasets $\mathbf{A} \in \mathbb{R}^{m\times n}$ while preserving its crucial information. This is achieved by transforming the original variables into a set of new, uncorrelated variables called principal components, which retain most of the variation from the original variables. The first principal component is the direction in feature space along which projections of observations have the largest variance. The second principal component is the direction which maximizes variance among all directions orthogonal to the first one, etc. In the first step, the data are centered by subtracting it mean vectors for each column from them. The covariance matrix $\mathbf{C} \in \mathbb{R}^{n\times n}$ for the columns (features) in matrix $\mathbf{B} = \left(\overrightarrow{b_1} \cdots \overrightarrow{b_n}\right)$ is calculated as $\mathbf{C} = \frac{1}{m}\mathbf{B}^T\mathbf{B}$, as well as the eigenvectors, and the corresponding eigenvalues, for matrix $\mathbf{C}$, such as $\mathbf{CW} = \mathbf{W\Lambda}$. Matrix $\mathbf{W}$ contains eigenvectors, and the diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues. For 2D plotting, we can project the data onto the first $k = 2$ components by truncating matrix $\mathbf{W}$ to k most significant features ($\mathbf{W}_k$) and projecting the data.

4. Solution overview

The key idea behind the proposed supply management flow (cf. Fig. 1) is to represent every print of sale (PoS) as a vector based on their historical sales patterns, grouping of similar points, and selecting the central (most representative) element. For instance, consider cluster C with central element PoS_C. Here the assumption is that without significant mismatch to customers' needs, we can supply all $\mathrm{PoS}_i \in C$ with the menu prepared for PoS_C, thereby significantly limiting the effort by $|C|$.

The whole process starts in the upper left corner in Fig. 1, with the extraction of two data representations and grouping of similar points of sale (cf. Section 4.1). Next, we apply DAAR reducts to discover the most discriminative features of clusters (cf. Section 4.2). This allows us to construct meaningful human-readable cluster descriptions, as presented in Section 4.3. In the following step, we engage an expert, whose tasks are limited to preparing delivery plans for the selected cluster prototypes (cf. Section 4.4). In this way, the overall effort is significantly limited. For instance, having 400 PoS in the pipeline and an average cluster size of 10, the expert's effort is limited to planning supply for just 40 PoS.

In the presented flowchart, the process may split depending on the human decision. In the main flow, the expert prepares the delivery for the selected PoS and triggers their automatic propagation cf. flow C in Fig. 1. In some situations, however, the provided cluster description may be considered ambiguous; here we may apply LLMs to refine the text (cf. Section 4.3). There is also a possibility to control the process on demand, cf. flow D. In particular, the developed solution has mechanisms for tracking changes in data by projecting them onto a two-dimensional space (cf. Section 4.5).

4.1. Data representation and clustering. Data ingestion and integration are two initial steps, leading to a vectorized representation of historical sales in every PoS. In our solution, we construct two alternative representations by aggregating transactions per individual product or their category: *prod_data* and *cat_data*, respectively (cf. Section 5.1). The product-based representation aggregates historical sales transactions by each product offered in a PoS. The category-based representation aggregates transactions for entire product categories. To compare these two representations, we rely on the L_1 (city-block) distance, which allows us to measure the dissimilarity between two vectors based on the sum of the absolute differences in their values.

In our experiments, we verified several clustering methods to group similar PoS (cf. Section 3.2). It was also vital to select the most representative PoS (prototype) for each cluster. The two flat clustering algorithms we used are *kmeans* and *pam* (Guo *et al.*, 2020). For *kmeans* the prototypes are chosen as the instances closest to the clusters' centers. For *pam*, we use medoids. For agglomerative clustering, we used *single_linkage*, *complete_linkage*, and *ward_linkage*. We used also the *di-*

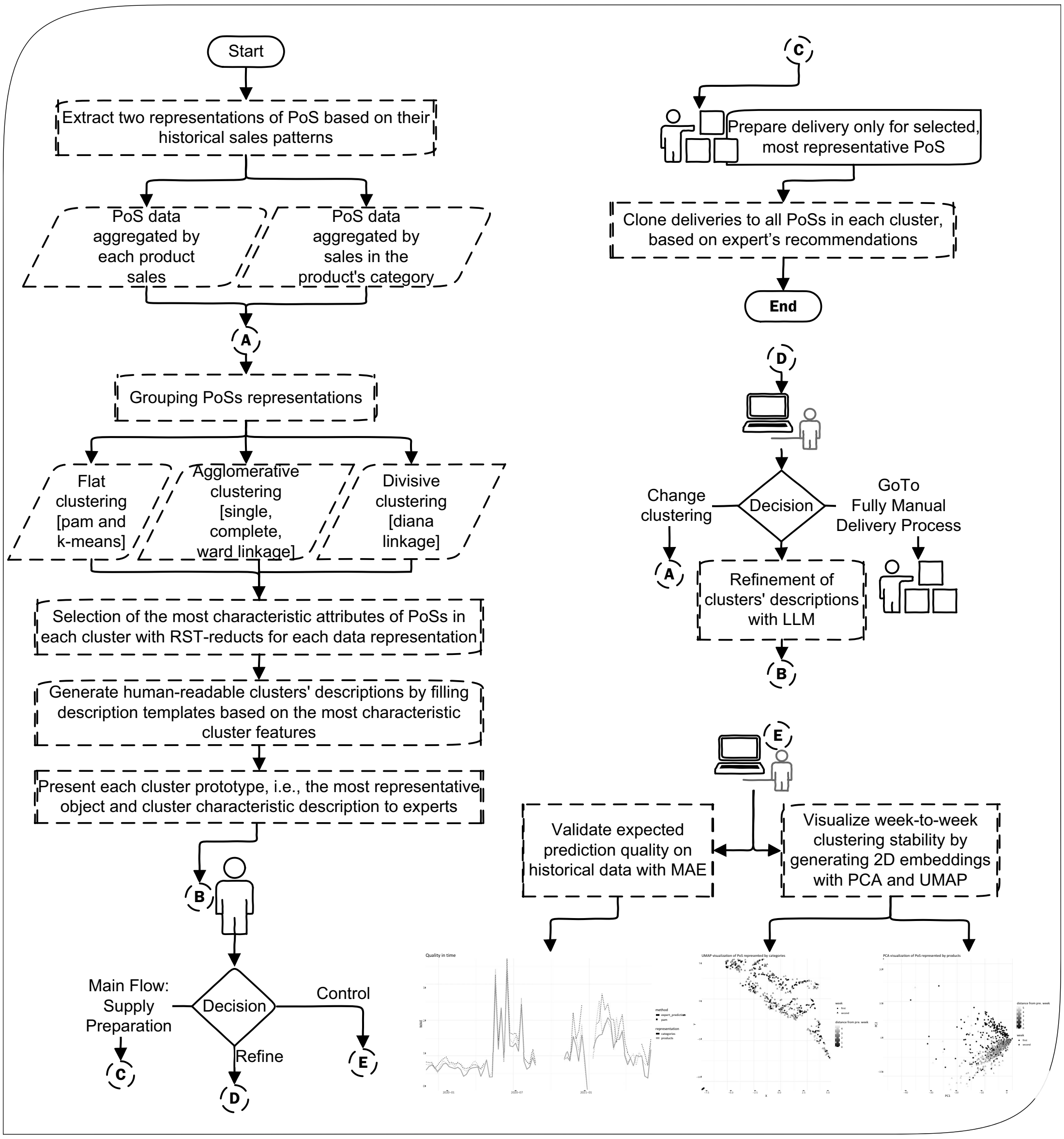
Start
Extract two representations of PoS based on their historical sales patterns
PoS data aggregated by each product sales
PoS data aggregated by sales in the product's category
A
Grouping PoSs representations
Flat clustering [pam and k-means]
Agglomerative clustering [single, complete, ward linkage]
Divisive clustering [diana linkage]
Selection of the most characteristic attributes of PoSs in each cluster with RST-reducts for each data representation
Generate human-readable clusters' descriptions by filling description templates based on the most characteristic cluster features
Present each cluster prototype, i.e., the most representative object and cluster characteristic description to experts
B
Main Flow: Supply Preparation
Decision
Control
Refine
C
D
E
C
Prepare delivery only for selected, most representative PoS
Clone deliveries to all PoSs in each cluster, based on expert's recommendations
End
D
Change clustering
Decision
GoTo Fully Manual Delivery Process
A
Refinement of clusters' descriptions with LLM
B
E
Validate expected prediction quality on historical data with MAE
Visualize week-to-week clustering stability by generating 2D embeddings with PCA and UMAP
Quality in time

Fig. 1. Flowchart of the proposed method.

ana_linkage method (Ezugwu *et al.*, 2022) which defines the cluster's diameter as the largest distance between any two members of a group. The details of each method are described in Section 3.2.

Additionally, we included *expert_prediction*, which is the top line for quality evaluation on prod_data, since it would correspond to singleton clusters and a random method (*random_random*) that shows the bottom line. To complete the picture, we also evaluate the importance of the selection of a prototype PoS that well represents a cluster, regardless of the clustering algorithm. In *random_custom*, the division into clusters is performed at random, yet the representative vending machine is selected as the closest to the cluster's center.

4.2. Attributes for cluster differentiation. A key aspect of the proposed framework is identifying the most distinctive attributes of objects belonging to different clusters. Our method leverages the concept of decision reduct, derived from rough set theory, which represents an irreducible set of attributes that sufficiently distinguishes objects in different decision classes. In our study, we discuss two approaches to reduct computation—a *global* approach that identifies a single set of attributes that is discriminative for all clusters, and a *local* approach that determines a unique set of attributes for each cluster, capturing only its specific characteristics.

In the *global* approach, we calculate a single decision reduct to discern PoS classes (here clusters). This is achieved using a greedy local discretization (Riza *et al.*,

Table 1. Characteristic attributes of exemplary clusters’ (global and local discernibility).

Method	Cluster		Cluster’s characteristic attributes	Attr. value		Ratio	
	No.	Size		High	Low	High	Low
Global	4	3	7 prev. days sales ‘Chicken Sechuan’	2	1	1.99	1.865
			7 prev. days sales ‘Tomato soup’	2	1	1.95	1.515
			sales in cat.‘Other meals’	19	11	1.459	1.323
			sales in cat.‘Pasta’	6	3	2.977	2.143
			sales in cat.‘Snacks’	3	1	1.206	3.32
Global	5	51	7 prev. days sales ‘Chicken Sechuan’	2	1	10.115	2.604
			7 prev. days sales ‘Tomato soup’	2	1	19.926	2.837
			sales in cat.‘Other meals’	19	11	14.907	2.205
			sales in cat.‘Pasta’	6	3	10.039	4.613
			sales in cat.‘Snacks’	3	1	6.084	1.472
Local	1	2	sales in cat.‘Other meals’	18	17	1.349	3.478
			total sales	54	45	3.181	1.497
Local	6	30	sales in cat.‘Other meals’	8.5	5	5.244	2.399
			total sales	56	38	3.54	1.01

2014), combined with the DAAR algorithm (Janusz and Ślęzak, 2015). The resulting reduct $R_{\text{DAAR}} \subseteq A$ provides a unified set of attributes that capture the variations in sales patterns across all clusters. In the *local* method, reducts are calculated individually for each cluster, but instead of discriminating all PoS from all clusters, they focus on a single cluster only (one vs. all). As a result, reducts are specialized to detect the most distinguishing factors of the corresponding group of PoS. They aid in identifying attribute cuts that result in rules with larger lift coefficients,[1] and thus lead to more meaningful cluster descriptions.

In the subsequent step, for each cluster $C \subseteq U$, attribute $a \in R_{\text{DAAR}}$, and its discretized value v, we estimate the *lift* of a rule $u \in C \implies a(u) = v$. Lastly, we construct natural language descriptions of clusters, highlighting their key characteristics. To ensure clarity, we may only include attributes with lift values exceeding a specified threshold, focusing on the most relevant cluster features. Table 1 presents sample results of both *global* and *local* reduction methods for data in the investigated case study (cf. Section 5). The global method results in high consistency in the description of all clusters because they are based on the same attributes. The difference is in the ratios that reflect the lift coefficients of the corresponding attribute value. Table 1 displays sample results of both global and local reduction methods used to build cluster descriptions for the case study under investigation. The global method yields longer descriptions, which, however, are highly consistent across all clusters, and distinctions among the PoS groups lie solely in the ratios that mirror the lift of the corresponding attribute value. The local method results in a more concise description, but the reducts may differ in their attributes, their discretization, and resulting lift values (cf. Table 1).

4.3. Cluster descriptions. It is important to keep cluster names possibly compact and easily interpretable by users. We achieve this through direct suggestions expressed by these names about what is in the clusters, which become particularly useful in reasoning conducted by intelligent systems regarding perceived situations. This is related to the very important problem of creating concepts in natural language.

To craft descriptions that emphasize the unique properties of objects within clusters, we apply the algorithms as described in Section 4.2. In this way, obtaining each cluster’s characteristics (cf. Table 1). We use pre-prepared formatted text templates to enhance the prepared descriptions and make them more self-descriptive. For the local approach, Cluster 6, and lift greater than 3.0, descriptions are “*PoS from Cluster* 6 *are* 5.244 *times more likely to have greater sales in* ***seven previous days*** *than* 8 *in the category* ***‘Other meals’*** *than other PoS*” and “*PoS from Cluster* 6 *are* 3.540 *times more likely to have greater* ***total sales*** *than* 56 *in* ***seven previous days*** *than other PoS*.”

For the global method, Cluster 5, and the lift threshold of 10.0, the generated cluster descriptions in the first row of Table 2. It is also important to mention that both methods are independent and may be computed simultaneously—it is up to the user to determine which is easier to understand. Additionally, as an optional step in the process, the user may call

[1] Lift is calculated by dividing the confidence of the rule $X \to Y$ by the support of the consequent (Y).

Table 2. Example descriptions for Cluster 5.

Method	Cluster description
Cluster 5 desc.	(1) PoS from Cluster 5 are 10.1 times more likely to have greater sales in *seven previous days* than 2 for *'Chicken Sechuan'* than other PoS (2) PoS from Cluster 5 are 19.9 times more likely to have greater sales in *seven previous days* than 2 for *'Tomato soup'* than other PoS (3) PoS from Cluster 5 are 14.9 times more likely to have greater sales in *seven previous days* than 19 in the category *'Other meals'* than other PoS, etc.
GPT-3.5	PoS in Cluster 5 have significantly higher sales in the last seven days compared with 2 other PoS for 'Chicken Sechuan', 'Tomato soup', and 19 PoS in the category of 'Other meals'. They are respectively 10.1 times, 19.9 times, and 14.9 times more likely to have greater sales in these categories compared to other PoS.
GPT-4	Points of Sales in Cluster 5 are significantly more likely to have high sales compared to other PoS. They are 10.1 times more likely to have high sales for 'Chicken Sechuan', 19.9 times more likely for 'Tomato Soup', and 14.9 times for 19 other meals in the previous week.
PaLM2	PoS from Cluster 5 are more likely to have increased sales of Chicken Sechuan and Tomato soup, and the chance of them having increased sales of other meals is also high.
Gemini Pro	Points of sale (PoS) in Cluster 5 are much more likely to have higher sales in several categories compared to other PoS: (1) They are over 10 times more likely to have higher sales of 'Chicken Sechuan' (2) They are almost 20 times more likely to have higher sales of 'Tomato soup' (3) They are nearly 15 times more likely to have higher sales in the 'Other meals' category This means that PoS in Cluster 5 are particularly strong performers in these categories and may have unique characteristics or strategies that contribute to their success.

a large language model (LLM) of choice (Min *et al.*, 2023). The prompt construction is based on a simple instruction that fits well for the autoregressive LLMs like GPT: *Summarize briefly Points of Sales in a cluster in a human-readable form, knowing that:* "⟨ORIGINAL DESCRIPTION⟩". Table 2 provides several examples of how selected LLMs can reformulate the initial prompt, i.e., ⟨ORIGINAL DESCRIPTION⟩ that may be found in the first row ('Cluster 5 desc.'). We may notice that the refinement of clusters' descriptions with LLM introduces a sort of vagueness and speculation into descriptions, e.g., *"... significantly more likely... "*, *"... chance of them having increased sales.."*, *"... are particularly strong performers in these categories... "*. In some cases, the level of creativity may be misleading (especially without numeric evidence), provide inaccuracies, or result in unclear statements like *"... strategies that contribute to their success... "*. In the future, we consider further evaluation of the usefulness of this approach by polling experts, or conducting Action Research (Przybyłek *et al.*, 2022).

4.4. Cluster prototypes. In order to best cater to customers' needs and provide attractive meal offerings, it would be ideal for experts to prepare the menu individually for each location. However, given a large number of locations, limited resources, and time, this would not be practical. The gist is to group points of sales based on the customers' purchase patterns (reflected in sales transactions) and then involve experts to prepare the menu for entire clusters of similar PoS. To achieve this, we first cluster similar PoS and then select the most representative one as a cluster prototype to depict the entire group. In this way, by maintaining an average cluster size of X, we ensure that the workload for experts is reduced by a factor of X. In our study, we aim for $X \geq 5$. We also assess whether this approach could yield satisfactory results through an exploration of real data sourced from `FitBoxY.com` (cf. Section 5).

4.5. Cluster stability visualization. The developed solution adjusts to variations in customer preferences. Still, such changes are reflected in the purchasing patterns, thereby causing the data representations to vary in time. Consequently, this affects clustering results, which may differ from week to week. We can observe these regularities by projecting vectors representing PoS onto a two-dimensional space and emphasizing the changes for every PoS. In our case, we use two significantly different approaches. The first one is UMAP that is strongly

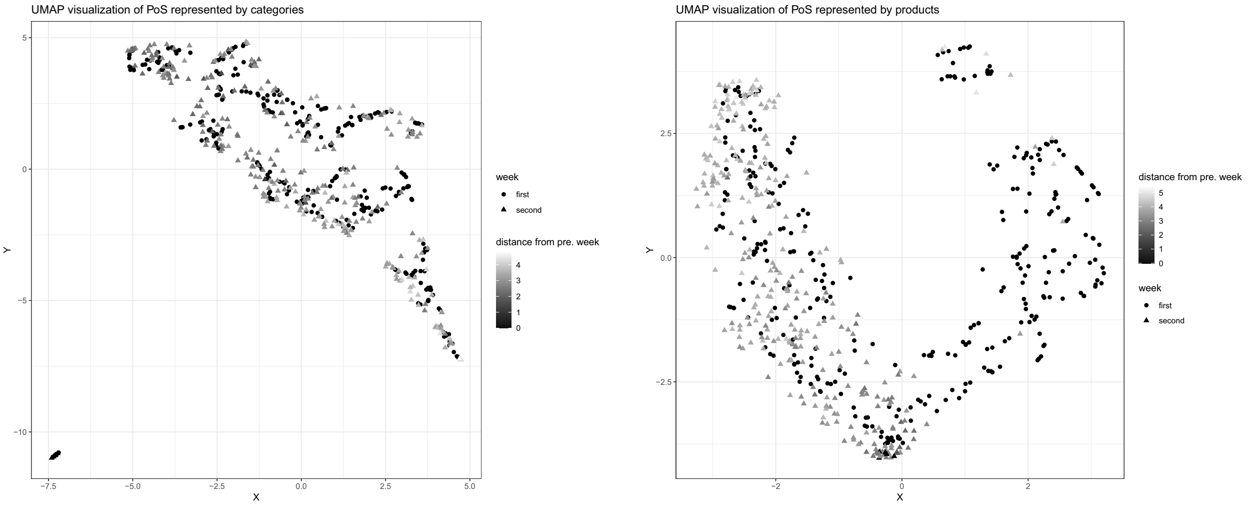

Fig. 2. 2D projection with UMAP that emphasizes the distance between the particular PoS in the first and second week.

nonlinear, whereas the other one is a linear dimensionality reduction technique, namely PCA. We provide two alternative visualizations because PCA, which may be easier to interpret, can also come short of capturing the similarities between instances when the first two principal components do not express a sufficient fraction of data variance.

Figure 2 shows 2D-embeddings with UMAP for *cat_data* and *prod_data* representations. Observably, the category-based representation is more stable, i.e., the points representing each PoS for two consecutive weeks in data form a more compact structure. We can also clearly see this regularity in Fig. 3 where we apply PCA as a dimensionality reduction method. In the figures, PoS from the first week are represented as circles, while those from the second week are represented as triangles. The distance between the same points from the first and second week is emphasized with the color depth – the further the city-block distance between vectors representing PoS for the second week from the first one, the brighter the color.

5. Experimental evaluation

5.1. Data. The data is sourced from `FitBoxY.com` vending machines and contains the sales history collected between June 21, 2017, and May 21, 2021. The time series data cover the first three waves of the COVID-19 pandemic with all the impacts caused by multiple lockdowns, as well as the more stable and uniform pre-pandemic period. The test data covers the period between December 2, 2019, and May 21, 2021, so they cover the first year of the pandemic and a period of a few months before the COVID-19 outbreak.

There are two versions of the dataset, aggregated by products and categories. The product-based dataset (*prod_data*) represents each point of sale (PoS) as a vector of all available products, indicating the quantity sold in the last week. The category-based dataset (*cat_data*) aggregates products into seven categories (e.g., breakfast, small lunch dishes, pasta). To capture temporal variations in the customer behavior, we considered each PoS at different time points as separate instances. For example, let A, B represent two PoS and t, t' two week-long periods in data, we have four different vectors in the data $A_t, A_{t'}, B_t, B_{t'}$. This approach allows us to model customer behavior changes over time at a given location.

5.2. Experiment flow. For both category and product data representations, we perform the following steps for each week in the test data set, i.e., ⟨2019-12-02, 2019-12-09⟩, ⟨2019-12-09, 2019-12-16⟩, . . . , ⟨2021-05-24, 2021-05-31⟩. We generate the training data involving all the instances before the beginning of the test week, as specified by the logic of the following query:

```
SELECT
  t.product, t.PoS, t.week_no, t.year,
  sum(t.product.qty) as sales_qty
FROM  transactions t
WHERE t.date < TEST_WEEK_START_DATE
GROUP BY
  t.product, t.PoS, t.week_no, t.year;
```

A similar query is used for *cat_data* with *t.product* being replaced by *t.product.category* (i.e., we additionally aggregate products into categories, hence achieving smaller and more dense vectors). Next, vectors of products and categories with aggregated ‘sales_qty’ values are created, which we subsequently cluster using the methods of choice (see Section 4.1) into clusters C. For each cluster C, we choose its most representative PoS (cf. Section 3.2). We similarly process the test data.

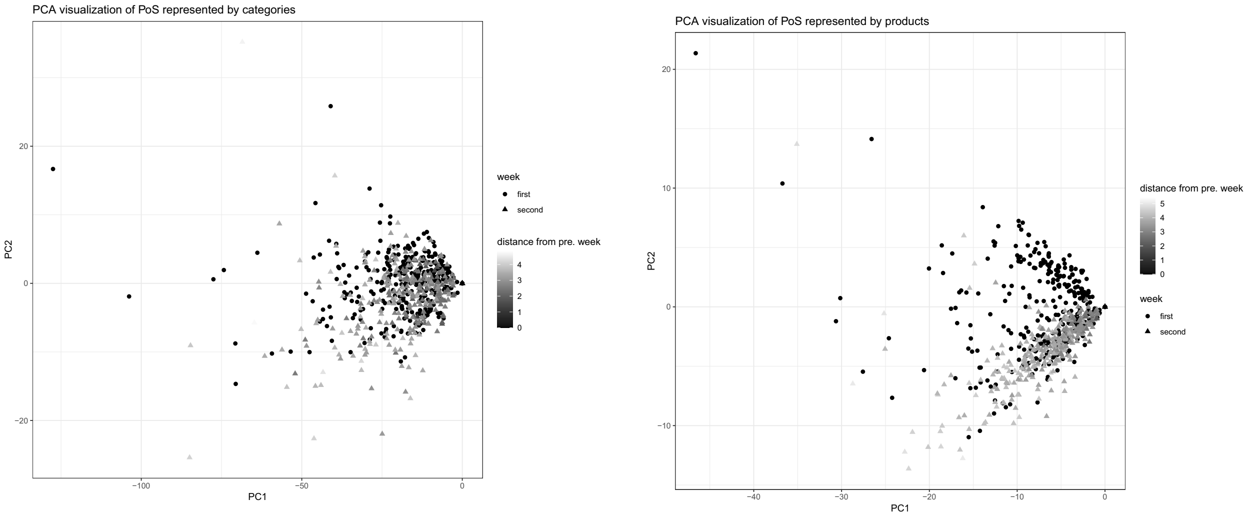

Fig. 3. 2D projection with PCA that emphasizes the distance between the particular PoS in the first and second week.

In the prediction phase, each PoS is allocated to the nearest cluster. For the delivery prediction for the test week, we consider products delivered to the most representative PoS of the corresponding cluster. These recommendations are then compared with the actual product sales at the PoS over the next seven days. We evaluate the quality of these predictions using the mean average error (MAE) and root mean square error (RMSE). Naturally, all the sales have already taken place (and are reflected in the collected data). For instance, suppose that based on the cluster C and its prototype PoS_C, we supply $\mathrm{PoS}_Y \in C$ with PoS_C menu. Proper clustering should assure high similarity between PoS_Y and PoS_C, but it may happen that some of the products available in PoS_C were not present in PoS_Y, and customers could not purchase such. Hence, this way of evaluation additionally penalizes the proposed approach, and in practice, we can expect even better results.

5.3. Experimental results. In an experimental evaluation, we compared the performance of the proposed framework depending on the data representation and clustering method of choice. In particular, we may notice that the PAM clustering algorithm achieved the best results for the product-based representation in terms of the mean absolute error (MAE) of 1.15, yet considering the RMSE slightly better results were achieved by Ward's minimum variance method (*ward_linkage*) (cf. Table 3). An interesting observation is that the random_custom method performed relatively well in terms of MAE. It could suggest that a reasonable selection of the most representative element is more important than the used clustering method. In terms of the RMSE, this method was less successful, which corresponds to a relatively high standard deviation of the results. This approach could work in a situation where sales from week to week are very stable and predictable, although when the data includes periods of high sales variability, the clustering method becomes important.

For *cat_data*, complete_linkage hierarchical clustering performed very well, with MAE of 1.26 and RMSE of 2.67 and the difference from the experts' predictions (which in practice correspond to singleton clusters) was not big. The detailed results of all methods on both data sets, including MAE, RMSE, and their standard deviations (*sd*) are presented in Table 3. Notably, we can achieve accurate demand estimations with approximately one-product mismatch in a week-long prediction horizon, yet significantly reducing the workload. These findings highlight the effectiveness of our proposed method for practical applications.

Figure 4 illustrates the fluctuation of error in time for the expert method and PAM clustering on both data representations. The empty sections correspond to periods of COVID-19 lockdowns. After the outbreak of the pandemic, as well as a few weeks before and after each lockdown, sales patterns deviated significantly from the historical mean, making such periods particularly difficult to predict. This resulted in noticeably higher prediction errors. However, as the pandemic progressed and more data was collected, models became better equipped to represent these unusual patterns. Consequently, the error levels observed towards the end of the chart are only marginally higher than those before the pandemic.

Furthermore, we investigated the relation between cluster size and our solution performance. We may notice the trend that the greater the number of elements in the cluster, the greater the error (both MAE and RMSE). For instance, in the case of the PAM algorithm and an average cluster size of approximately 12, the MAE error is 0.86, whereas for clusters of size 15 MAE is 1.06,

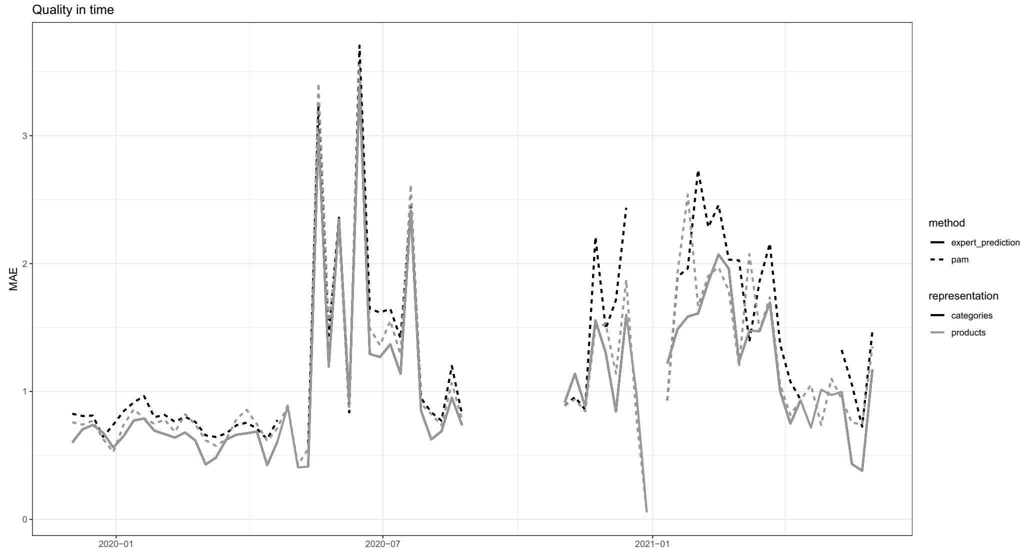

Fig. 4. MAE in time for two data representations for *pam* and *expert_prediction*.

for clusters of approximately 20 elements MAE is 1.21. We may notice that this regularity was less visible before the pandemic when the difference in MEA between small and big clusters lies between 0.72 and 0.75. This may be related to the similar performance of the vending machines and consistent menus, which somehow explains the good performance of the random_custom method. The differences are more significant during the pandemic, where depending on the cluster size MEA varies between 1.03 and 1.46. This period is, however, particularly hard for analysis because the number of test data points varies in time, i.e., there are periods with significantly less operating PoS and it is impossible to construct bigger clusters.

In terms of evaluating the stability of clustering over time (cf. Section 4.5), representations based on categories are generally more stable. Given that the sum of all attribute values in both representations is identical, it is possible to compare their distributions of week-to-week distances with the L_1 (city-block) distance. The city-block distance between the points of sale represented by categories is relatively small, with a mean of 18.85, a median of 15, and a maximum of 123. These values are noticeably lower compared with the product-based representation, which has a mean of 41.49, a median of 37, and a maximum of 225. Typically, the overall dispersion of PoS for two consecutive weeks is similar for the first representation, whereas it is decidedly different for the second one (cf. Figs. 2 and 3). Our experiments demonstrated that for an average cluster size of approximately 5 vending machines, the product-based representation (*prod_data*) yields better results than the category-based representation (*cat_data*), yet the latter one is much more concise and behave more stable.

6. Summary

This study discusses some essential challenges related to trustworthiness and interpretability in ML applications and presents a novel approach to supply management in the FMCG market that leverages rough sets, clustering, and dimensionality reduction to enable human-computer interaction. In the proposed approach, instead of demand forecasting with predictive or prescriptive machine learning techniques, we directly rely on experts' recommendations that are scaled to hundreds of objects (here, points of sales) utilizing rough sets (RST) and unsupervised clustering algorithms and their central points. The solution is focused not only on prediction accuracy but also on stability in time and interpretability. In the presented case study of the FMCG market, we showed that it is possible to operate on whole groups of PoS, significantly reducing the work required to prepare delivery plans resulting in increased work efficiency. The conducted experimental evaluation confirmed that the proposed approach achieved a fair trade-off between ML performance, interpretability, and trustworthiness.

One of the future research directions and improvements is related to interaction with experts (Grzegorowski, 2023). Following the idea of Lofti Zadeh, assuming that *"information granulation plays a key role in the implementation of the strategy of divide-and-conquer in human problem-solving,"* one can consider using this idea to decompose the specification of the problem considered in the work expressed in natural language (Averkin, 2023). The decomposition of such specification can be carried out in dialogue with users, supported by knowledge discovered from data about the system's operation, e.g., regarding the location of

Table 3. MAE and RMSE of demand prediction (*prod_data* vs. *cat_data*).

	Method	*prod_data*				*cat_data*			
		MAE	sd	RMSE	sd	MAE	sd	RMSE	sd
1	expert_prediction	1.05	0.60	2.45	1.46	1.07	0.62	2.50	1.50
2	random_random	1.25	0.72	2.66	1.55	1.35	0.74	2.81	1.61
3	random_custom	1.16	0.7	2.55	1.59	1.35	0.73	2.77	1.60
4	kmeans	1.17	0.66	2.54	1.53	1.27	0.71	2.68	1.54
5	pam	1.15	0.66	2.53	1.54	1.3	0.72	2.72	1.54
6	single_linkage	1.16	0.67	2.53	1.54	1.41	0.74	2.82	1.47
7	complete_linkage	1.18	0.67-	2.55	1.53	1.26	0.67	2.67	1.51
8	ward_linkage	1.17	0.65	2.53	1.52	1.28	0.68	2.73	1.51
9	diana_linkage	1.17	0.67	2.53	1.54	1.29	0.69	2.71	1.53

PoS, variability in the sales over time, preferences of users in the vicinity of PoS, etc. Another possibility is to explore the applicability of pre-trained language models and template-based fine-tuning (Min *et al.*, 2023). Another important aspect of future research is the further development of adaptive strategies for the developed system. It would also be valuable to manage the behavior of the modeled system through distributed control implemented in a network of interacting local models.

Acknowledgment

Research co-funded by the Polish National Centre for Research and Development (NCBiR), grant no. POIR.01.01.01-00-0963/19-00, and by the Polish National Science Centre (NCN), grant no. 2018/31/N/ST6/00610.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* **6**: 52138–52160, DOI: 10.1109/ACCESS.2018.2870052.

Averkin, A. (2023). Ideas of Lotfi Zadeh in explainable artificial intelligence, *in* S.N. Shahbazova *et al.* (Eds), *Recent Developments and the New Directions of Research, Foundations, and Applications*, Springer, Cham, pp. 45–48.

Barredo Arrieta, A., Diaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* **58**: 82–115, DOI: 10.1016/j.inffus.2019.12.012.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. and Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions, *ACM Computing Surveys* **55**(9): 1–33, DOI: 10.1145/3561048.

Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I. and Akinyelu, A.A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects, *Engineering Applications of Artificial Intelligence* **110**: 104743, DOI: 10.1016/j.engappai.2022.104743.

Fisher, A., Rudin, C. and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, *Journal of Machine Learning Research* **20**(177):1–81.

Grzegorowski, M. (2023). Selected aspects of interactive feature extraction, *in* J.F. Peters *et al.* (Eds), *Transactions on Rough Sets XXIII*, Springer, Berlin/Heidelberg, pp. 121–287, DOI: 10.1007/978-3-662-66544-2_8.

Grzegorowski, M., Janusz, A., Lazewski, S., Swiechowski, M. and Jankowska, M. (2022). Prescriptive analytics for optimization of FMCG delivery plans, *in* D. Ciucci *et al.* (Eds), *Proceedings of IPMU'22*, Springer, Berlin/Heidelberg, pp. 44–53.

Grzegorowski, M., Janusz, A., Śliwa, G., Marcinowski, L. and Skowron, A. (2023). Towards ML explainability with rough sets, clustering, and dimensionality reduction, *in* A. Campagner *et al.* (Eds), *Proceedings of IJCRS 2023*, Springer, Berlin/Heidelberg, pp. 371–386.

Grzegorowski, M. and Ślęzak, D. (2019). On resilient feature selection: Computational foundations of r-C-reducts, *Information Sciences* **499**: 25–44, DOI: 10.1016/j.ins.2019.05.041.

Guo, X., Lin, H., Wu, Y. and Peng, M. (2020). A new data clustering strategy for enhancing mutual privacy in healthcare IoT systems, *Future Generation Computer Systems* **113**: 407–417, DOI: 10.1016/j.future.2020.07.023.

Habbal, A., Ali, M.K. and Abuzaraida, M.A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions, *Expert Systems with Applications* **240**: 122442, DOI: 10.1016/j.eswa.2023.122442.

Heide, N.F., Muller, E., Petereit, J. and Heizmann, M. (2021). X^3SEG: Model-agnostic explanations for the semantic segmentation of 3D point clouds with prototypes and criticism, *2021 IEEE International Conference on Image Processing (ICIP), Anchorage, USA*, pp. 3687–3691.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Jia, H. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Information Sciences* **622**: 178–210, DOI: 10.1016/j.ins.2022.11.139.

Janusz, A. and Ślęzak, D. (2015). Computation of approximate reducts with dynamically adjusted approximation threshold, *in* F. Esposito *et al.* (Eds), *Proceedings of ISMIS 2015*, Springer, Berlin/Heidelberg, pp. 19–28.

Kannout, E., Grzegorowski, M., Grodzki, M. and Nguyen, H.S. (2024). Clustering-based frequent pattern mining framework for solving cold-start problem in recommender systems, *IEEE Access* **12**: 13678–13698.

Malefors, C., Secondi, L., Marchetti, S. and Eriksson, M. (2021). Food waste reduction and economic savings in times of crisis: The potential of machine learning methods to plan guest attendance in Swedish public catering during the COVID-19 pandemic, *Socio-Economic Planning Sciences* **82**(A): 101041.

Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I. and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey, *ACM Computing Surveys* **56**(2): 1–40, DOI: 10.1145/3605943.

Pawlak, Z. (1982). Rough sets, *International Journal of Computer and Information Sciences* **11**: 341–356.

Pawlak, Z. and Skowron, A. (2007). Rudiments of rough sets, *Information Sciences* **177**(1): 3–27.

Penta, A. and Pal, A. (2021). What is this cluster about? explaining textual clusters by extracting relevant keywords, *Knowledge-Based Systems* **229**: 107342.

Pięta, P. and Szmuc, T. (2021). Applications of rough sets in big data analysis: An overview, *International Journal of Applied Mathematics and Computer Science* **31**(4): 659–683, DOI: 10.34768/amcs-2021-0046.

Przybyłek, A., Albecka, M., Springer, O. and Kowalski, W. (2022). Game-based sprint retrospectives: Multiple action research, *Empirical Software Engineering* **27**(1): 1.

Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślęzak, D. and Benítez, J.M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package 'RoughSets', *Information Sciences* **287**(0): 68–89.

Stawicki, S., Ślęzak, D., Janusz, A. and Widz, S. (2017). Decision bireducts and decision reducts—A comparison, *International Journal of Approximate Reasoning* **84**: 75–109.

Tarallo, E., Akabane, G.K., Shimabukuro, C.I., Mello, J. and Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer goods: An exploratory research, *IFAC-PapersOnLine* **52**(13): 737–742.

Zhang, C.-X., Zhang, J.-S. and Yin, Q.-Y. (2017). A ranking-based strategy to prune variable selection ensembles, *Knowledge-Based Systems* **125**: 13–25.

Zong, W., Chow, Y. and Susilo, W. (2020). Interactive three-dimensional visualization of network intrusion detection data for machine learning, *Future Generation Computer Systems* **102**: 292–306.

Marek Grzegorowski has long been associated with the Institute of Informatics at the University of Warsaw, where, in 2021, he received his PhD in computer science. He is an active and experienced researcher in the fields of science related to data exploration, machine learning, and artificial intelligence. In his career, he has conducted several R&D projects related to the application of ML/AI in academic and industry collaboration.

Andrzej Janusz is an active academic scientist and experienced researcher in fields related to data exploration, machine learning, and artificial intelligence. In 2014, he received his PhD in computer science from the University of Warsaw, where he held the position of an assistant professor. Since 2024, he has been with the Queensland University of Technology. His research projects have been related to safety monitoring in hazardous environments, video game analytics, active learning, and explainable AI. He is a co-founder of `KnowledgePit.ai`—an online data science platform, where he organizes international competitions.

Łukasz Marcinowski is a co-founder and a board member of the company FitFood Poland responsible for logistics and distribution. He is a successful entrepreneur who strongly believes in innovation, finding practical applications of modern technologies in business. He supports several initiatives to improve operational excellence in the FMCG industry by adopting machine learning and prescriptive analytics.

Andrzej Skowron is an ECCAI (EurAI), AAIA and IRSS fellow, a member of Academia Europaea, an EU Academy of Sciences fellow, and a WI Academy founding fellow. He is a full professor in the Systems Research Institute, Polish Academy of Sciences, and a professor emeritus of the Faculty of Mathematics, Computer Science and Mechanics at the University of Warsaw. His areas of expertise include approximate reasoning, rough sets, (interactive) granular computing, intelligent systems, (adaptive) complex systems, perception-based computing, and machine learning.

Dominik Ślęzak received his PhD in computer science in 2002 from the University of Warsaw, where he currently works. He had also worked at the Polish-Japanese Academy of Information Technology and the University of Regina. In 2020, was awarded the professorial title by the President of Poland. He has co-authored over 200 articles in the fields of data mining, databases, and rough sets, and co-invented over 20 US patents. He has also chaired over 20 scientific conferences. He is the vice-president of the Polish AI Society.

Grzegorz Śliwa is a co-founder and CEO of the company FitFood Poland. He is an entrepreneur with a firm belief in the importance of innovation and applying technologies in practical business aspects. He champions various initiatives aimed at enhancing operational excellence within the consumer packaged goods industry through harnessing data and the adoption of machine learning and advanced analytics.

Received: 25 March 2024
Revised: 31 July 2024
Accepted: 9 January 2025