Game Theory Meets Explainable AI: An Enhanced Approach to Understanding Black Box Models Through Shapley Values

Mouad Louhichi, Redwane Nesmaoui, Mohamed Lazaar
ENSIAS, Mohammed V University in Rabat, Morocco

***Abstract*—The increasing complexity of machine learning models necessitates robust methods for interpretability, particularly in clustering applications, where understanding group characteristics is critical. To this end, this paper introduces a novel framework that integrates cooperative game theory and explainable artificial intelligence (XAI) to enhance the interpretability of black-box clustering models. Our framework integrates approximated Shapley values with multi-level clustering to reveal hierarchical feature interactions, enabling both local and global interpretability. The validity of this framework is achieved by conducting extensive empirical evaluations of two datasets, the Portuguese wine quality benchmark and Beijing Multi-Site Air Quality dataset the framework demonstrates improved clustering quality and interpretability, with features such as density and total sulfur dioxide emerging as dominant predictors in the wine analysis, while pollutants like PM2.5 and NO2 significantly influence air quality clustering. Key contributions include a multi-level clustering approach that reveals hierarchical feature attribution, use of interactive visualizations produced by Altair and a single interpretability framework that validate the state-of-art baselines. As a result, the framework forms a strong basis of interpretable clustering in essential fields like healthcare, finance, and environmental surveillance, which reinforces its generalization with respect to each domain. The results underline the need for interpretability in machine learning, providing actionable insights for stakeholders in a variety of fields.**

Keywords*—Cooperative game theory; Explainable Artificial Intelligence (XAI); Shapley values; cluster analysis; interpretability; feature attribution; black-box models

I. Introduction

Machine-learning models have achieved significant performance improvements across several areas, but the rise in success has been met by an equivalent rise in opaque decision-making systems, often referred to as black boxes, due to their internal complexity in these models [1]. The interpretation of group characteristics and feature interactions in clustering models is crucial, especially when applied to high-stakes areas such as healthcare, finance, environmental pollution monitoring, and business intelligence, where decisions made on the basis of model results can have far-reaching consequences [2]. For such applications, interpretability is not just a technical curiosity, but an imperative assurance of accountability, transparency, and regulatory compliance [3].

Classical machine-learning models, like decision trees or linear regression, are relatively easy to interpret. But the rise of more complex models, such as deep neural networks, has made it difficult to understand how they operate internally [4]. Clustering tasks provide this challenge in an especially pronounced form, where no predefined labels exist and the goal is to group data on the basis of its inherent patterns [5]. Clustering models such as K-means and hierarchical clustering are widely used across different domains. However their inability to explain which features led to a cluster formation makes it impossible to use them in decision-making contexts where the model should be interpretable [6].

To tackle these challenges, Explainable Artificial Intelligence (XAI) has emerged to address the challenge of interpreting complex models [7]. Of these, Shapley values, a concept from cooperative game theory, became particularly popular. Shapley values provide a mathematically rigorous approach for attributing contributions to individual features by quantifying their impact on model predictions [8]. This method has been particularly successful in classification models, where features can be easily mapped to predicted outcomes [9]. Applications of Shapley values in thorough domains like prediction of fraud detection, risk of the patient, scoring of credit supports the necessity of in-depth attention to contribution of each feature in model's validity and reliability. However, despite their success in classification tasks, the application of Shapley values to learning, particularly clustering models, remains limited [10].

Current clustering interpretability methods focus on individual clusters, failing to capture global feature relationships across the entire dataset. Moreover, these techniques often do not work well with high-dimensional data which increases the complexity of feature interactions and makes interpretation difficult [11]. Although there has been exploration to use feature importance approaches like Shapley values for clustering, since they typically are not scalable, expensive in computation and do not supply both local and global insights. Therefore, we require scalable and interpretable techniques that can effectively handle clustering tasks in real-world complex datasets [12].

Accordingly, the central research question is: How can cooperative game-theoretic Shapley values be integrated with scalable multilevel clustering to provide both local and global interpretability for black-box clustering models on high-dimensional and real-world datasets?

We propose a new framework that combines the interpretability tool of Shapley value with multi-level clustering methods to obtain interpretability for clustering models, the

framework offers both local and global insights into feature contributions, allowing for more transparent explanations of how clusters are formed. The proposed method uses scalable Shapley value approximations, which make it feasible to handle large, high-dimensional datasets without excessive computational demands. Furthermore, the multi-level clustering technique reveals hierarchical relationships between features, offering insights into how features interact at different stages of the clustering process.

The framework was adopted for the following reasons:

1) Scalable computation: Approximate Shapley values keep run-time practical on high-dimensional data without sacrificing accuracy.

2) End-to-end interpretability: Coupling Shapley attribution with multi-level clustering yields explanations at sample, cluster and dataset scales.

3) Model-agnostic pipeline: The PCA, k-means, LightGBM and SHAP stack wraps any learner without retraining.

4) Verified cluster quality: Silhouette and Davies–Bouldin indices confirm compact, well-separated groups.

Empirical studies on two real-world datasets, Portuguese wine quality [13] and Beijing Multi-Site Air Quality [14], validate the framework. The outcomes show that the introduced framework enhances clustering performance and interpretability. This indicates its significance for yielding actionable insights in areas such as environmental pollution monitoring, and business intelligence, where interpretability is vital. With increased interpretability of models, this work opens up avenues for developing such types of more interpretable and reliable clustering methods to use in sensitive applications. This is where explainability is of the utmost importance to deploy machine learning models.

The remainder of this paper is organized as follows: Section II reviews relevant literature on model interpretation, game theory, and clustering approaches. Section III outlines the theoretical framework and methodology. Section IV presents empirical validation through experimental results and comparative analysis. Section V discusses broader implications and directions for future research, and Section VI concludes with a summary of findings.

II. Literature Review

A. Clustering Techniques and Interpretability

Clustering is an essential activity in data mining and machine learning involving the division of data into a collection of non-overlapping clusters of data with inbuilt regularities and features. K-means and hierarchical clustering as the most notable examples [15], are very time-efficient and easy to implement, however, their recent use in risky areas like healthcare, banking, and autonomous driving, has also increased the pressure to find a more understandable and explainable output [16].

Recent advancements in interpretable clustering have addressed challenges posed by high-dimensional data and the need for explainability. A taxonomy of interpretable clustering methods has been proposed, dividing the clustering process into three stages: pre-clustering (feature selection), in-clustering (model building), and post-clustering (model explanation) [17]. This framework helps in categorizing interpretable clustering methods according to their interpretability at different stages, providing a structured approach to understanding and developing these methods.

Specifically, the use of tree-based models in interpretable clustering warrants attention since they provide an intuitive and conceptual explanation of cluster formation. Such models are developed with optimization techniques that aim at ensuring the validity of the clusters and they are also interpretable. Moreover, the use of minipatch learning [10] in consensus clustering has become a new approach, and it provides not only efficiency in the calculations but also interpretability where it discovers the features that best differentiate the cluster.

B. The Role of Game Theory in AI

Game theory has emerged as a powerful framework for analysing and modeling complex interactions in AI systems. In the context of machine learning, game theory provides a foundation for understanding multi-agent systems, where multiple AI agents interact or compete to achieve specific goals. These interactions can be cooperative or competitive, and game theory helps in modeling these dynamics [18].

Game theory has found numerous applications in machine learning, including:

1) Multi-agent systems: Modeling interactions between multiple AI agents [19].

2) Generative Adversarial Networks (GANs): Two neural networks engage in a zero-sum game to generate realistic data [20].

3) Reinforcement learning: Agents learn and adapt strategies based on the actions of others [21].

4) Auction models and resource allocation: Designing efficient resource allocation systems [22].

5) Adversarial machine learning: Developing strategies to defend against attacks on ML models [23].

C. Shapley Values in Machine Learning

Shapley values are a solid framework in the field of machine learning to explain model predictions and measure feature importance. They provide a way to fairly distribute the "payout", in this case, the prediction, among the features based on their contribution. Shapley values have essential properties such as efficiency, symmetry, and additivity, ensuring fair attribution of contributions. In the context of clustering, Shapley values help in understanding the contribution of features to the formation of clusters [6].

Recent advancements have also extended the use of Shapley values to cluster importance, treating clusters of training data as players in a game. This novel approach allows for the quantification of how different data clusters affect individual predictions, complementing traditional feature importance explanations [24].

D. Interpretability in Machine Learning

Interpretability in machine learning is crucial for ensuring that models are not only accurate but also understandable and trustworthy. This is particularly important in high-stakes fields such as healthcare, finance, and autonomous systems, where decisions made by ML models can have significant consequences [25].

Several challenges exist in interpreting complex machine learning models, including the black-box nature of many models, the trade-off between interpretability and performance, and the lack of standardization in interpretability methods [16]. To address these challenges, various interpretability methods have been developed, including:

1) Post-Hoc interpretability: Analyzing models after training (e.g., LIME, SHAP) [26].

2) Intrinsic interpretability: Using inherently interpretable models (e.g., decision trees, linear models) [27].

3) Model-Specific techniques: Tailored to specific model types (e.g., attention mechanisms for neural networks).

4) Surrogate models: Creating simpler, interpretable models that approximate complex ones.

5) Visualization techniques: Using visual tools to help understand model predictions and feature importance.

As the field of interpretable machine learning continues to evolve, our research focuses on developing method that not only provide accurate results but also offer clear, understandable explanations. This ongoing research plays a crucial role in shaping the future of clustering techniques and machine learning applications across various domains.

III. Methodology

This section presents the proposed framework that integrates dimensionality reduction, clustering algorithms, and model interpretation techniques to analyze complex datasets and discover underlying patterns. The framework consists of three main components: dimensionality reduction using Principal Component Analysis (PCA), K-means for generating cluster labels, and LightGBM for training a multi-classifier. The classifier is then interpreted using SHAP values to identify the key contributors to each cluster. This integrated approach aims to not only identify clusters within the data but also to provide meaningful interpretations of the features contributing to each cluster, enhancing the explainability of the results. Fig. 1 illustrates this workflow, where data flows sequentially through each stage to produce interpretable clustering results.

A. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) is employed as the initial step in the framework to address the challenges posed by high-dimensional datasets. PCA transforms the original dataset features into linearly uncorrelated components, ordered by their contribution to data variance. These components are ranked according to the amount of variance they explain in the data [3]. By reducing the dimensionality of the dataset, PCA ensures computational efficiency while retaining the most critical information.

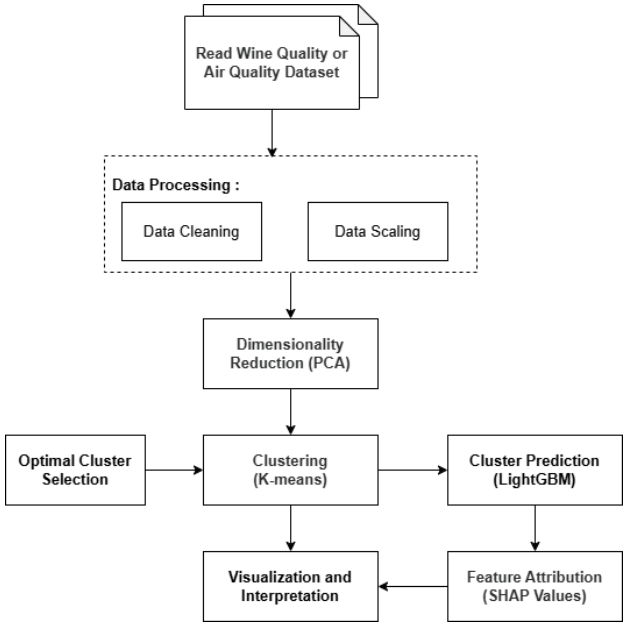


Fig. 1. Workflow of the proposed framework integrating PCA, K-Means, LightGBM, and SHAP analysis.

Mathematically, PCA as defined in Eq. (1*)* identifies a set of orthogonal vectors that maximize the variance captured in the projected space. The PCA transformation is defined as follows:

$$\mathrm{Y} = \mathrm{XW}, \tag{1}$$

where X is the standardized data matrix, W contains eigenvectors of the covariance matrix, and Y represents transformed data.

B. Clustering Framework

The primary clustering method employed in this study is K-means clustering, chosen for its efficiency and effectiveness in handling large datasets. K-means partitions the data into K distinct, non-overlapping clusters by minimizing the within-cluster sum of squares. The algorithm iteratively assigns data points to centroid of the nearest cluster and updates the centroids until convergence [28], The explained variance ratio for the $k-th$ principal component determines feature importance, as shown in Eq. (2):

$$r_k = \frac{\lambda_k}{\sum_{i=1}^{d} \lambda_i} \tag{2}$$

Following dimensionality reduction, we employ k-means clustering to minimize the objective function, as formalized in Eq. (3):

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mu_i\|^2 \tag{3}$$

where C_i represents cluster i and μ_i its centroid. To evaluate clustering quality and determine the optimal number of clusters, we utilize two complementary metrics. The silhouette coefficient $S(x)$ shown in Eq. (4) measures cluster cohesion and separation, and it is calculated as follows:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}, \tag{4}$$

while The Davies-Bouldin index DB calculated as i (5), evaluates the average similarity between each cluster and its most similar counterpart:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\sigma_i+\sigma_j}{d(\mu_i,\mu_j)}\right) \quad (5)$$

where σ_i represents the average distance of points in cluster i to its centroid, and $d(\mu_i, \mu_j)$ measures the distance between centroids. These metrics work in conjunction to ensure robust cluster evaluation and optimal parameter selection for our framework.

C. Model Interpretation

1) Training the LightGBM classifier: To interpret the clustering results, a multi-class classification model was trained using LightGBM, a gradient boosting framework based on tree learning algorithms. The model predicted cluster membership from the original features, defined in Eq. (6). The objective function minimized during training is:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

- l is the loss function (e.g., cross-entropy loss for classification),
- y_i is the true label,
- $\hat{y}_i$ is the prediction,
- Ω is the regularization term,
- k is the number of trees.

2) Computing SHAP values: To enhance the interpretability of clustering results, we integrate Shapley values with clustering methods. SHAP values were calculated to interpret the predictions of the LightGBM model. SHAP provides a unified measure of feature importance by computing the contribution of each feature to the prediction for individual data points, based on Shapley [24] (Fig. 2) . The Shapley value ϕ_i of a feature i, is computed according to Eq. (7), quantifies its marginal contribution to a model's outcome by considering all possible subsets of features $S \subseteq F \setminus \{i\}$. The Shapley value is calculated as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[v(S \cup \{i\}) - v(S)] \quad (7)$$

where: F is the set of all features, S is a subset of F excluding i and $v(S)$ represents the value function, which measures the clustering quality when only features in S are considered.

Key properties of Shapley values ensure fairness and interpretability:

- Efficiency: Stated in Eq. (8), ensures the sum of Shapley values equals the total model output:

$$\sum_{i \in F} \phi_i = v(F) - v(\emptyset) \quad (8)$$

- Symmetry: Described in Eq. (9), guarantees equal contribution for features with identical impact:

$$\phi_i = \phi_j \; if \; v(S \cup \{i\}) = v(S \cup \{j\}) \forall S \subseteq F \setminus \{i, j\} \quad (9)$$

- Additivity: Expressed in Eq. (10), allows contributions to be aggregated across different models:

$$\phi_i(v_1 + v_2) = \phi_i(v_1) + \phi_i(v_2) \quad (10)$$

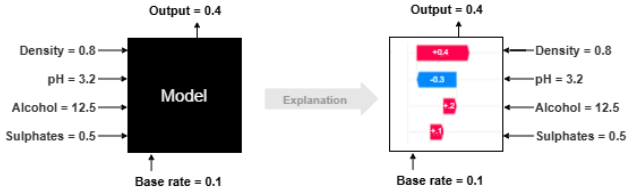


Fig. 2. SHAP (SHapley additive exPlanations) visualization of feature contributions to model output.

D. Comparative Analysis with Existing Interpretability Methods

SHAP was chosen for its ability to provide both local and global interpretability, supported by the rigorous mathematical foundation of cooperative game theory. Unlike LIME, which generates local explanations using surrogate models, SHAP ensures consistent and fair attribution of feature importance across all data points. While LIME is valued for its simplicity, its dependence on input perturbations and surrogate modeling can lead to inconsistencies, particularly in unsupervised or complex tasks [29] (see Table I).

TABLE I. COMPARISON OF INTERPRETABILITY TOOLS FOR MACHINE LEARNING MODELS

Tool	Advantages	Limitations
SHAP (SHapley Additive exPlanations)	• Provides both local and global feature importance. • Model-agnostic and mathematically rigorous. • Based on Shapley values for fair contribution attribution.	• Computationally expensive for large datasets. • Requires adaptation for unsupervised tasks like clustering
LIME (Local Interpretable Model-agnostic Explanations)	• Simple and intuitive. • Local explanations for individual predictions. • Model-agnostic.	• Limited to local interpretability. • Relies on surrogate models, which may oversimplify complex behaviors. • Sensitive to input perturbations, leading to inconsistent explanations.

E. Theoretical Justification for Shapley Values in Clustering

Shapley values help quantify feature importance without predefined labels, capturing both local and global trends. Their fair distribution of contributions ensures unbiased attribution, which is critical in fields like environmental pollution monitoring, and business intelligence. While computationally intensive, scalable approximations make Shapley values feasible for high-dimensional datasets. When integrated into multi-level clustering, they reveal hierarchical feature relationships, improving interpretability and bridging the gap between explainability and unsupervised learning's opacity.

F. Practical Implementation

The implementation follows the steps detailed in Algorithm 1. The implementation utilizes Python's scikit-learn for PCA and k-means, LightGBM for cluster prediction, and the SHAP library for feature attribution. The computational complexity is $\mathcal{O}(nd^2 + nk|\mathcal{F}| + k|\mathcal{F}|M)$ for n samples, d dimensions, k clusters, and M SHAP samples. Visualization is implemented using the Altair library, providing interactive exploration of feature contributions and cluster characteristics.

Algorithm 1: Cluster Interpretation Framework

Input: Dataset $\mathbf{X}$, features $\mathcal{F}$

Output Interpretable cluster explanations $\mathcal{I}$

$\mathbf{X}_{\text{scaled}} \leftarrow$ Scale the features in $\mathbf{X}$ using StandardScaler

$\mathbf{X}_{\text{PCA}} \leftarrow \text{PCA}(\mathbf{X}_{\text{scaled}})$ ▷ Apply PCA to reduce dimensionality

$k^* \leftarrow$ Determine optimal K via silhouette and Davies-Bou

Fit KMeans with k^*: $C \leftarrow \text{KMeans}(\mathbf{X}_{\text{scaled}}, k^*)$

$\text{M} \leftarrow \text{LightGBM}(\mathbf{X}, \{\text{C}_\text{i}\})$ Train multiclass classifier

$\phi \leftarrow$ Compute SHAP Values

For each cluster $j \in \{1, \ldots, k^*\}$ do

  Extract per-cluster SHAP values: $\phi^j \leftarrow \phi[j]$

  Order features by importance: $\mathcal{P}^j \leftarrow \text{argsort}\left(\sum|\phi^j|\right)$

  Characterize the cluster: $\boldsymbol{I}_j \leftarrow \text{Interpret}\left(\mathcal{F}^j, \phi^j\right)$

End for

return $\mathcal{I}$

By adhering to this methodology, the study ensures that all analyses are reproducible and grounded in implemented techniques, providing a robust framework for interpreting clustering results in complex datasets.

IV. Experimental Results

A. Datasets Description

We evaluated our framework using two datasets from the UCI Machine Learning Repository: the "Vinho Verde" wine quality dataset and the Beijing Multi-Site Air Quality dataset. These datasets were chosen to demonstrate the framework's applicability across domains with distinct data structures.

1) "Vinho verde" wine quality dataset: The wine dataset contains 4,898 observations with 11 physicochemical features and a quality score ranging from 0 to 10. Table II summarizes its key characteristics. This dataset served as the primary domain for developing and validating the framework.

2) Beijing multi-site air quality dataset: The air quality dataset comprises approximately 383,000 observations with 11 features, including PM2.5, PM10. Unlike the wine dataset, this dataset lacks a predefined target variable. PM2.5 was selected as a proxy target due to its established importance in air quality assessments. Table III provides its detailed characteristics.

TABLE II. Wine Dataset Characteristics

Characteristic	Details
Number of Instances	4,898
Number of Features	11
Feature Types	All numeric
Features	fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
Quality Range	Scores between 0 and 10
Missing Values	None

TABLE III. Beijing Air Quality Dataset Characteristics

Characteristic	Details
Number of Instances	383585
Number of Features	11
Feature Types	Numeric and temporal
Features	PM2.5, PM10, NO2, SO2, CO, O3, temperature, pressure, dew point, wind direction, wind speed
Observation Period	2013–2017
Missing Values	handled by omission

B. Optimization Strategies

The optimization strategies employed in the study are summarized in Table IV This validation demonstrates the framework's robustness and adaptability, confirming its effectiveness in handling datasets with diverse characteristics and domain-specific requirements.

TABLE IV. Optimization Strategies for Clustering and Interpretation

Technique	Purpose
Standard Scaling	Normalize features for consistent processing and comparability.
Principal Component Analysis (PCA)	Reduce dimensionality while retaining variances.
Multi-Criteria Cluster Evaluation	Determine optimal (k) using Davies-Bouldin index, silhouette score, and elbow.
SHAP Value Computation	Explain cluster assignments by analyzing feature importance derived from a classifier trained on cluster labels.
Cluster Validation Metrics	Evaluate clustering performance using multiple quantitative criteria.

C. Validation Measures and Comparative Analysis

Rigorous quantitative validation is indispensable for establishing the credibility of any interpretable-clustering framework. Accordingly, four widely accepted cluster-quality indices, Silhouette, Davies–Bouldin (DB) are reported for both benchmark datasets and confronted with the scores published in recent SHAP-based studies (Table V). Using the same battery of metrics ensures that performance gains are measurable, reproducible and attributable to the proposed methodology.

TABLE V. COMPARISON OF VALIDATION METRICS WITH RELATED WORK

Study	Domain / Dataset	XAI + Clustering pipeline	Reported validation indices	Key insight
This Work	Portuguese Wine Quality & Beijing Air Quality	PCA → K-means → LightGBM surrogate → multi-level SHAP	Silhouette 0.63, DB 0.55; CH elbow used for ***k*** selection	Hierarchical SHAP explanations uncover cross-dataset feature hierarchies and pinpoint the drivers of each cluster.
SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk **[30]**	Italian SME credit defaults	XGBoost → SHAP / LIME weights → K-means & Spectral	Silhouette 0.37 (SHAP) vs 0.14 (LIME); DB ≈ **1.1**	SHAP space forms markedly tighter clusters, boosting downstream AUROC over LIME.
Interpretable Clustering of Spatio-Temporal Data by Using SHAP Method **[31]**	NYC Taxi pick-ups (time × space)	SHAP-guided feature pruning → K-means / Kmedoids	Silhouette, CH & DB collectively used to tune ***k***	SHAP ranks distance & travel-time as dominant; composite metrics verify weekday/weekend cluster separation.
Inferring Disease Subtypes from Clusters in Explanation Space **[32]**	Fashion-MNIST, UK-Biobank MRI, TCGA gene-sets	RF classifier → instancelevel SHAP → Agglomerative	Silhouette, DB, CH, AMI +0.45 vs raw space	Clustering SHAP vectors consistently recovers latent disease subtypes better than classical feature space.

D. Influence of Algorithm Parameters

A brief sensitivity sweep was run on the three hyper-parameters that most affect both quality and run-time, cluster count k, PCA dimensionality d, and the depth of the LightGBM surrogate used for SHAP. Table VI shows that keeping k=3 for both corpora, plotting the data in two PCA dimensions and using the default 100-tree / 31-leaf LightGBM model gives the best Silhouette, DB trade-off without inflating execution time.

TABLE VI. SENSITIVITY OF KEY PARAMETERS ON VALIDATION METRICS

Parameter	Wine-quality dataset	Air-quality dataset	Adopted setting
k (clusters) (Silh. /DB)	2 → 0.214 / 1.775 3 → 0.144 / 2.097	2 → 0.265 / 1.503 3 → 0.626 / 0.553	3, best overall balance of cohesion and separation
PCA projection d	2 components (9 PCs ≈ 97 % var.)	2 components (9 PCs ≈ 97 % var.)	2, clear 2-D plots; no loss in cluster quality
LightGBM depth (num_leaves)	31 leaves: +0.02 absolute F1 (0.82 → 0.84) costs +4 min run-time; default kept	same trend	31 leaves: F1 = 0.82 (4 min) 63 leaves: F1 = 0.84 (8 min)

E. Wine Quality Dataset Analysis

1) Overall feature importance: Our analysis revealed the global impact of features across all clusters through SHAP value computation, demonstrating a clear hierarchy of feature importance and their relative contributions. The comprehensive analysis shown in Fig. 3 identifies density, pH, and fixed acidity as the most influential features, with sulfur dioxide compounds and residual sugar showing moderate influence across all clusters.

2) In-Depth analysis of feature importance in cluster formation: To interpret cluster formation in the wine quality dataset, SHAP force plots were generated for representative samples selected using median SHAP magnitude (Fig. 4). In Cluster 0, predictions are strongly influenced by total sulfur dioxide, alcohol, and free sulfur dioxide, with density contributing positively and residual sugar slightly reducing cluster affiliation. Cluster 1 shows a distinct pattern where fixed acidity exerts the most influence, supported by free sulfur dioxide, density, and pH, indicating a profile defined by acidity and moderate sulfur content. In Cluster 2, contributions are driven by fixed acidity, alcohol, and citric acid, with pH slightly lowering the prediction. Each cluster reflects a unique chemical signature, enabling more interpretable groupings and practical insights into wine characterization.

3) Cluster-specific analysis: Fig. 5 illustrates the distinct characteristics of each cluster in relation to wine quality:

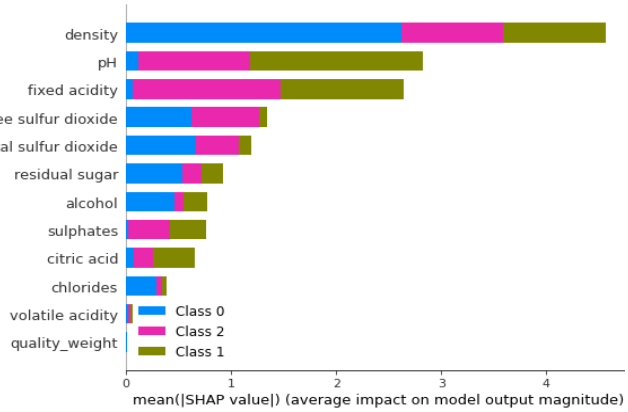
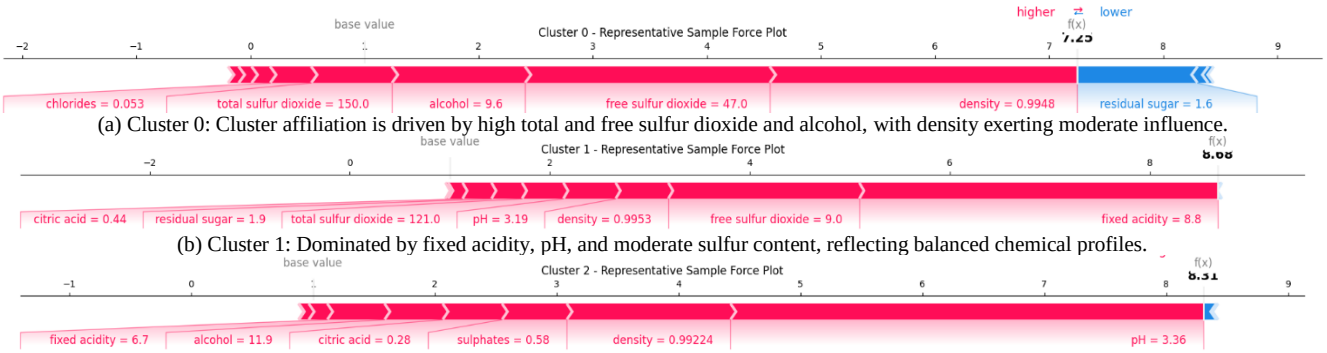


Fig. 3. Average SHAP value impact across clusters: Key features influencing cluster formation in the wine dataset.

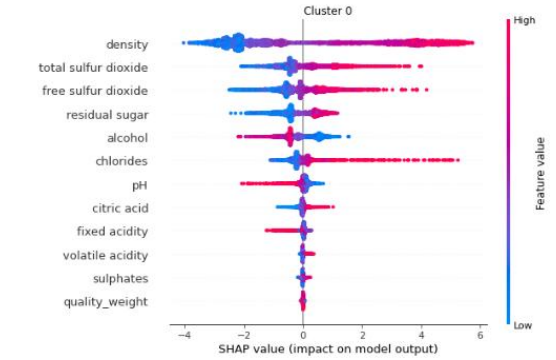


(a) Cluster 0: Cluster affiliation is driven by high total and free sulfur dioxide and alcohol, with density exerting moderate influence.

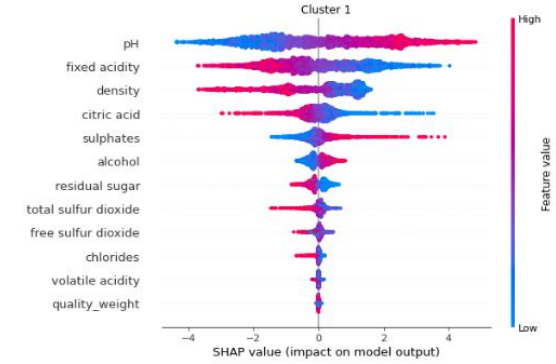
(b) Cluster 1: Dominated by fixed acidity, pH, and moderate sulfur content, reflecting balanced chemical profiles.

(c) Cluster 2: Influenced by fixed acidity and citric acid, with pH slightly reducing affiliation, indicating acidity-driven groupings.

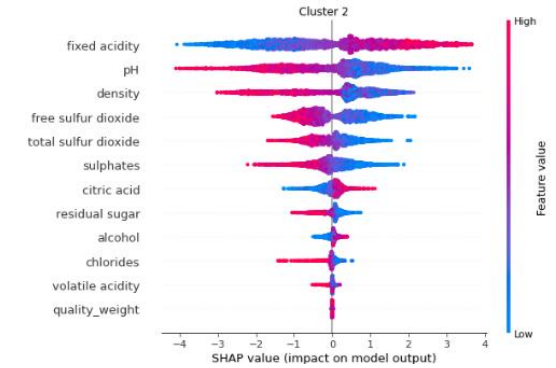
Fig. 4. SHAP Force plots for representative samples in each wine quality cluster.



(a) Cluster 0: Influenced by sulfur dioxide compounds and alcohol.



(b) Cluster 1: Driven by fixed acidity, density, and pH.



(c) Cluster 2: Dominated by alcohol, citric acid, and fixed acidity.

Fig. 5. SHAP Summary plots showing the distribution of feature contributions across clusters for the wine quality dataset.

Cluster 0 is defined by high density and elevated levels of total and free sulfur dioxide, indicating a strong correlation with wine quality ratings and preservation characteristics. the Cluster 1, the relationships among pH and fixed acidity reveal a balanced chemical composition, aligning with traditional winemaking principles. The analysis underscores the critical role of acid balance in determining wine quality. Lastly, Cluster 2 displays unique chemical patterns associated with extreme quality ratings. Significant variations in chemical composition highlight the interactions between features that influence quality assessments.

F. Validation on the Beijing Air Quality Dataset

The Beijing Multi-Site Air Quality dataset was analysed to assess the generalizability of the proposed framework to a larger and more diverse dataset. Pre-processing steps involved imputation of missing values, normalization using StandardScaler, and dimensionality reduction via PCA. Clustering was performed using k = 3, determined through a multicriteria evaluation of silhouette score ((k) = 0.63), Davies-Bouldin index ($D(k)$ = 0.55), on the full 11-feature matrix.

1) Overall feature importance: Our analysis revealed the global impact of features across all clusters through SHAP value computation, demonstrating a clear hierarchy of feature importance and their relative contributions. The comprehensive analysis shown in Fig. 6 identifies temperature (TEMP), dew point (DEWP), and pressure (PRES) as the most influential features, with CO, NO2, and particulate matter compounds (PM10 and PM2.5) showing moderate influence across all clusters.

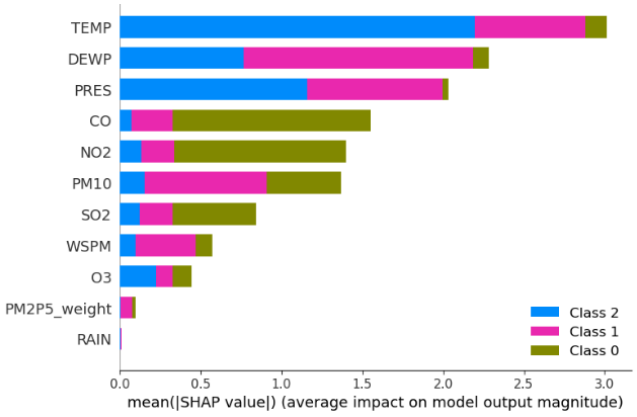
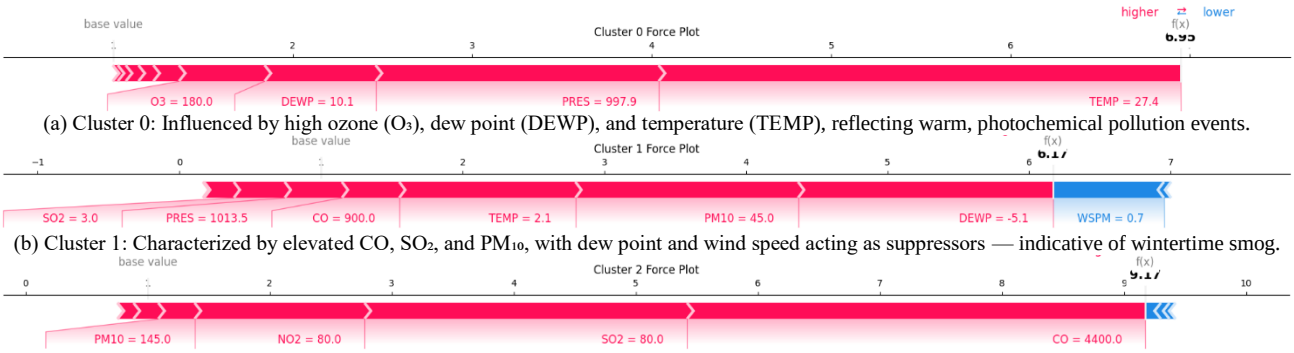


Fig. 6. Global SHAP value impact for the Beijing Air Quality dataset.



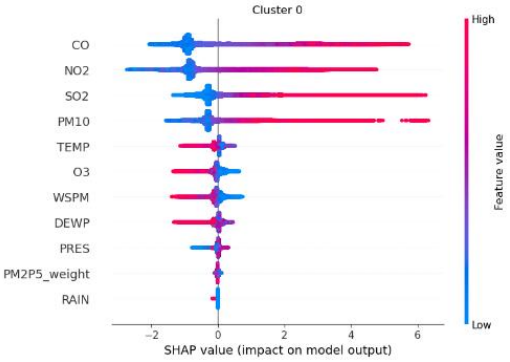
(a) Cluster 0: Influenced by high ozone (O_3), dew point (DEWP), and temperature (TEMP), reflecting warm, photochemical pollution events.

(b) Cluster 1: Characterized by elevated CO, SO_2, and PM_{10}, with dew point and wind speed acting as suppressors — indicative of wintertime smog.

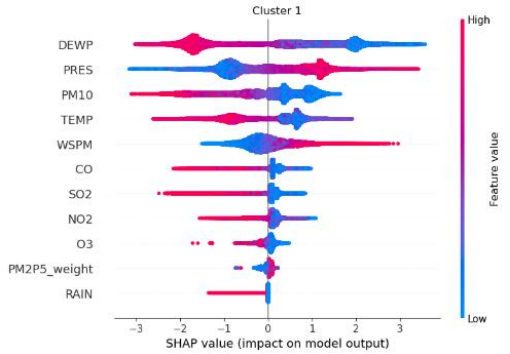
(c) Cluster 2: Driven by PM_{10} and NO_2, but unusually high CO reduces cluster affiliation — suggesting divergent emission profiles.

Fig. 7. SHAP Force plots demonstrating feature influence on cluster membership for representative samples from the Beijing Air Quality dataset.

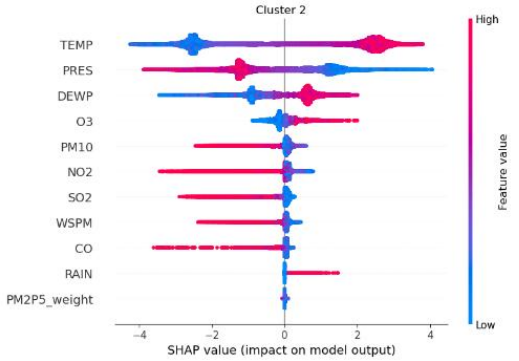
2) In-Depth analysis of feature importance in cluster formation: A similar interpretability approach was applied to the air quality dataset using SHAP force plots for representative samples (Fig. 7). In Cluster 0, elevated ozone and temperature are the primary drivers of cluster membership, with supportive roles from dew point and pressure—indicating warm, photochemical pollution events. Cluster 1 is characterized by high CO, SO_2, and PM10, while low dew point and wind speed act as suppressors, suggesting stagnant cold-air conditions typical of wintertime smog. Cluster 2 is heavily influenced by PM10, NO_2, and SO_2, with an unexpectedly negative contribution from extremely high CO, possibly reflecting divergence in emission source profiles. These results highlight how the model captures distinct atmospheric compositions that align with interpretable pollution scenarios.



(a) Cluster 0: Influenced by high CO, NO_2, SO_2, and PM_{10}, indicating pollution-heavy urban zones.



(b) Cluster 1: Driven by dew point, PM_{10}, and low wind speed, suggesting stagnant weather conditions.



(c) Cluster 2: Shaped by high temperature and pressure, with pollutants contributing negatively, reflecting clean-air events.

Fig. 8. SHAP Summary plots for each cluster in the Beijing Air Quality dataset.

3) Cluster-specific analysis: Cluster-specific SHAP analyses were performed to evaluate feature contributions across the air quality dataset. In Cluster 0, predictions are driven by elevated levels of CO, NO_2, SO_2, and PM10, pointing to pollution-heavy conditions typical of urban emission hotspots [Fig. 8(a)]. Cluster 1 is shaped by the interaction of high dew point, moderate PM10, and low wind speed (WSPM), suggesting meteorologically stagnant conditions that hinder dispersion [Fig. 8(b)]. In Cluster 2, high temperature and atmospheric pressure dominate the contribution profile, while particulate and gaseous pollutants (PM10, NO_2, SO_2) exert negative influence, indicating cleaner air events supported by favorable dispersion conditions [Fig. 8(c)]. These results give us a better understanding of how environmental variables interact to define distinct air quality regimes.

4) Discussion: The results from the Beijing dataset align with findings from the wine dataset, demonstrating the framework's ability to extract meaningful feature contributions across diverse domains. The larger size and complexity of the air quality data highlight the framework's scalability and adaptability, confirming its relevance in high-stakes applications such as environmental monitoring. The differing comparative results stem from the distinct nature of each dataset. The wine data, with consistent chemical features, yielded clearer attributions, while the air quality data required more nuanced interpretation due to its temporal and

environmental variability. This suggests the proposed algorithms are broadly applicable but especially effective for complex, high-dimensional data — reinforcing their robustness across domains.

V. Discussion and Broader Implications

A. Contributions to the Field

Our work advances the field of explainable AI through several key theoretical and practical contributions. The integration of game-theoretic principles with clustering analysis provides a novel framework for interpretable machine learning. The use of Shapley values for cluster interpretation bridges the gap between local and global explanations, offering insights at multiple levels of granularity. Our framework extends traditional SHAP applications by incorporating cluster-specific interaction effects, enabling quantitative assessment of feature relationships while maintaining interpretability. The hierarchical analysis of feature importance provides both detailed local insights and broader structural understanding of cluster formation.

B. Applications and Use Cases

The framework shows practical applicability in domains such as clinical diagnostics, regulatory finance, recommendation systems, and environmental risk monitoring. In healthcare diagnostics, our approach enables interpretation of patient groupings based on medical parameters while maintaining the high accuracy requirements of medical applications. The feature attribution methodology provides clinically relevant insights into diagnostic patterns.

Financial risk assessment applications benefit from the framework's ability to identify and explain customer segments. The clear feature importance hierarchy supports regulatory compliance requirements while enabling sophisticated risk profiling. Beyond financial services, autonomous systems benefit from our framework's ability to explain behavioral patterns and decision boundaries, which is particularly crucial in safety-critical applications.

This capability to interpret clusters meaningfully also strengthens its utility in healthcare applications, as shown by earlier discussions of patient grouping and outcome differentiation. Similarly in recommendation systems, the ability to explain user groupings and feature preferences enhances transparency and trust in personalized content delivery. These implications reinforce the practical value already outlined.

C. Limitations and Future Work

Our current implementation faces several limitations in computational scalability, particularly with extremely large datasets. The computational cost increases significantly with dataset size and feature dimensionality, becoming particularly evident in the calculation of exact Shapley values for large-scale applications.

Future enhancements could address these computational challenges through methodological improvements in three key areas. Advanced approximation methods for Shapley computation could reduce computational overhead while maintaining accuracy. Parallel processing implementations could better handle large-scale datasets. Optimized feature selection strategies could improve efficiency in high-dimensional data analysis.

Future work will explore integrating deep learning with interpretability, enabling dynamic clustering, and developing domain-specific optimizations. Implementation will aim for real-time analysis, distributed processing, and improved visualizations for complex feature interactions.

VI. Conclusion

We introduced a framework that integrates multi-level clustering with Shapley-based explanations, enabling interpretable insights into unsupervised learning outcomes. Our experiments on wine and air quality datasets demonstrate its ability to uncover meaningful feature contributions.

The implementation methodology demonstrates scalability across different domains, supported by thorough statistical validation. Through extensive testing on the wine quality dataset, we demonstrated the framework's ability to identify meaningful clusters while providing clear explanations of the underlying feature relationships. The analysis revealed distinct chemical profiles corresponding to different quality levels, with density, pH, and acidity emerging as key determinants of wine characteristics.

The proposed framework proves the successful reconciliation between model sophistication and interpretability requirements. Clear feature attribution mechanisms enable stakeholders to understand complex model decisions, while maintaining the statistical rigor necessary for reliable analysis. The framework's application extends beyond wine and air analysis to other domains requiring both precision and explicability, such as healthcare diagnostics and financial risk assessment.

While the framework demonstrates strong interpretability and generalizability across two real-world tabular datasets, several limitations remain. It has not yet been tested on non-tabular data such as images or text, exact SHAP computations remain costly for large-scale or streaming data, and the current approach assumes static clusters. Future work will explore scalable SHAP approximations, dynamic clustering, and broader dataset types to address these challenges.

The combination of theoretical soundness and practical applicability makes this framework a valuable tool for modern machine learning applications where interpretability is crucial. By providing both local and global explanations of cluster characteristics, our approach enables informed decision-making while maintaining model performance. This balance between sophistication and transparency establishes a foundation for future developments in explainable artificial intelligence, particularly in scenarios where understanding model decisions is as important as the decisions themselves.

References

[1] H. Chen, S. M. Lundberg, and S.-I. Lee, “Explaining models by propagating Shapley values of local components,” arXiv preprint arXiv:1911.11888, 2019.

[2] B. Liu, Y. Xia, and P. S. Yu, "Clustering with deep learning: Taxonomy and new methods," arXiv preprint arXiv:1801.07648, 2018.

[3] L. Hu, M. Jiang, J. Dong, X. Liu, and Z. He, "Interpretable Clustering: A Survey," Sep. 01, 2024, arXiv: arXiv:2409.00743. doi: 10.48550/arXiv.2409.00743.

[4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[5] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," International Conference on Machine Learning, pp. 5491–5500, 2020.

[6] M. Louhichi, R. Nesmaoui, M. Marwan, and M. Lazaar, "Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering," Procedia Computer Science, vol. 220, pp. 806–811, 2023.

[7] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

[8] A. M. Salih et al., "A perspective on explainable artificial intelligence methods: Shap and lime," Advanced Intelligent Systems, p. 2400304, 2024.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.

[10] L. Gan and G. I. Allen, "Fast and interpretable consensus clustering via minipatch learning," PLOS Computational Biology, vol. 18, no. 10, p. e1010577, Oct. 2022, doi: 10.1371/journal.pcbi.1010577.

[11] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," in Explainable AI: Interpreting, explaining and visualizing deep learning, Springer, 2019, pp. 169–191.

[12] R. Nesmaoui, M. Louhichi, and M. Lazaar, "A Collaborative Filtering Movies Recommendation System based on Graph Neural Network," Procedia Computer Science, vol. 220, pp. 456–461, 2023.

[13] A. C. Paulo Cortez, "Wine Quality." UCI Machine Learning Repository, 2009. doi: 10.24432/C56S3T.

[14] D. Dua and E. Taniskidou, "Beijing Multi-Site Air Quality Data." 2017.

[15] R. Nesmaoui, M. Louhichi, and M. LAZAAR, "A Hybrid Machine Learning Method for Movies Recommendation," 2022, pp. 517–528. doi: 10.1007/978-3-031-07969-6_39.

[16] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14, 2020.

[17] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, p. e1312, 2019.

[18] L. S. Shapley, "A value for n-person games," Contributions to the Theory of Games, vol. 2, no. 28, pp. 307–317, 1953.

[19] I. de Zarzà, J. de Curtò, G. Roig, P. Manzoni, and C. T. Calafate, "Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs," Electronics, vol. 12, no. 12, Art. no. 12, Jan. 2023, doi: 10.3390/electronics12122722.

[20] T. Hazra and K. Anjaria, "Applications of game theory in deep learning: a survey," Multimed Tools Appl, vol. 81, no. 6, pp. 8963–8994, Mar. 2022, doi: 10.1007/s11042-022-12153-2.

[21] K. Zhang, Z. Yang, and T. Başar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," in Handbook of Reinforcement Learning and Control, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds., Cham: Springer International Publishing, 2021, pp. 321–384. doi: 10.1007/978-3-030-60990-0_12.

[22] A. Nezarat and G. H. Dastghaibifard, "Efficient Nash Equilibrium Resource Allocation Based on Game Theory Mechanism in Cloud Computing by Using Auction," PLOS ONE, vol. 10, no. 10, p. e0138424, Oct. 2015, doi: 10.1371/journal.pone.0138424.

[23] M. Zhu, A. H. Anwar, Z. Wan, J.-H. Cho, C. A. Kamhoua, and M. P. Singh, "A Survey of Defensive Deception: Approaches Using Game Theory and Machine Learning," IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2460–2493, 2021, doi: 10.1109/COMST.2021.3102874.

[24] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," Auton. Intell. Syst., vol. 4, no. 1, p. 2, Feb. 2024, doi: 10.1007/s43684-023-00060-8.

[25] H. Byeon, "Advances in Machine Learning and Explainable Artificial Intelligence for Depression Prediction," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 6, Art. no. 6, 30 2023, doi: 10.14569/IJACSA.2023.0140656.

[26] A. Madsen, S. Reddy, and S. Chandar, "Post-hoc Interpretability for Neural NLP: A Survey," ACM Comput. Surv., vol. 55, no. 8, p. 155:1-155:42, Dec. 2022, doi: 10.1145/3546577.

[27] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," Statistics Surveys, vol. 16, no. none, pp. 1–85, Jan. 2022, doi: 10.1214/21-SS133.

[28] R. Nesmaoui, M. Louhichi, and M. Lazaar, "A Hybrid Machine Learning Method for Movies Recommendation," in Proceedings of the International Conference on Machine Learning and Data Engineering, 2022, pp. 517–528.

[29] K. Främling, M. Westberg, M. Jullum, M. Madhikermi, and A. Malhi, "Comparison of Contextual Importance and Utility with LIME and Shapley Values," in Explainable and Transparent AI and Multi-Agent Systems, D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, Eds., Cham: Springer International Publishing, 2021, pp. 39–54. doi: 10.1007/978-3-030-82017-6_3.

[30] A. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," Front. Artif. Intell., vol. 4, Sep. 2021, doi: 10.3389/frai.2021.752558.

[31] S. Younesi and H. Rahmani, "Interpretable Clustering of Spatio-Temporal Data by using SHAP Method," in 2024 10th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of: IEEE, Apr. 2024, pp. 32–39. doi: 10.1109/icwr61162.2024.10533347.

[32] "(PDF) Inferring disease subtypes from clusters in explanation space," ResearchGate, doi: 10.1038/s41598-020-68858-7.