applied sciences

MDPI

Article

Refining Filter Global Feature Weighting for Fully Unsupervised Clustering

Fabian Galis [1,*], Darian M. Onchis [1,*] and Codruta Istin [2]

[1] Department of Computer Science, West University of Timisoara, 300223 Timișoara, Romania
[2] Department of Computers, Politehnica University of Timisoara, 300006 Timișoara, Romania; codruta.istin@upt.ro
* Correspondence: fabian.galis@e-uvt.ro (F.G.); darian.onchis@e-uvt.ro (D.M.O.)

Featured Application

Our novel feature weighting strategies offer a practical tool for improving the clustering accuracy of unlabeled data, relevant in domains like bioinformatics, marketing segmentation, and anomaly detection.

Abstract

In the context of unsupervised learning, effective clustering plays a vital role in revealing patterns and insights from unlabeled data. However, the success of clustering algorithms often depends on the relevance and contribution of features, which can differ between various datasets. This paper explores feature weighting for clustering and presents new weighting strategies, including methods based on SHAP (SHapley Additive exPlanations), a technique commonly used for providing explainability in various supervised machine learning tasks. By taking advantage of SHAP values in a way other than just to gain explainability, we use them to weight features and ultimately improve the clustering process itself in unsupervised scenarios. Our empirical evaluations across five benchmark datasets and clustering methods demonstrate that feature weighting based on SHAP can enhance unsupervised clustering quality, achieving up to a 22.69% improvement over other weighting methods (from 0.586 to 0.719 in terms of the Adjusted Rand Index). Additionally, these situations where the weighted data boosts the results are highlighted and thoroughly explored, offering insight for practical applications.

Keywords: explainable AI; SHAP; unsupervised learning; feature weighting; clustering

check for updates

Academic Editor: Giorgio De Nunzio

Received: 26 June 2025
Revised: 8 August 2025
Accepted: 14 August 2025
Published: 18 August 2025

Citation: Galis, F.; Onchis, D.M.; Istin, C. Refining Filter Global Feature Weighting for Fully Unsupervised Clustering. *Appl. Sci.* **2025**, *15*, 9072. https://doi.org/10.3390/app15169072

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Clustering is a fundamental task in unsupervised learning that aims to group unlabeled data into meaningful subgroups (clusters) based on similarity measures. It has been applied extensively in a broad range of fields such as image segmentation, customer profiling, and bioinformatics, serving as a primary technique to discover hidden structures in data without relying on labels or annotations [1,2]. Numerous clustering algorithms have been developed; some of the most well-known include k-means, hierarchical clustering (e.g., Ward's method), and density-based algorithms such as DBSCAN [3,4]. Despite their widespread use and versatility, the effectiveness of clustering algorithms often depends heavily on the selected feature representation and dataset, which can substantially influence how similarities or distances among data points are measured [1,2].

Feature weighting (FW) has become an effective approach to mitigating the impact of irrelevant or less informative features on clustering algorithms. Rather than assigning equal importance to all features, global FW methods allocate specific weights to features according to their relevance to the clustering objective. Generally, FW techniques can be categorized into two main types based on their strategies for estimating these weights:

1. Filter. FW methods determine weights based on the relationship between the features and a specified reference, which corresponds, in an unsupervised scenario, to the intrinsic characteristics of the data [5].
2. Wrapper. FW methods utilize feedback from a given machine learning algorithm to estimate weights in an iterative, black-box manner. Based on the performance achieved in the previous iteration, which is calculated using either supervised or unsupervised evaluation metrics, the method determines whether to adjust the weights to enhance the model's performance for the next iteration or not [5].

In addition to FW offering a degree of interpretability regarding feature importance, dedicated XAI (explainable artificial intelligence) techniques have been developed for this purpose. For example, SHAP (SHapley Additive Explanations) decomposes individual predictions into feature-level contributions by utilizing concepts derived from cooperative game theory [6,7]. SHAP has primarily been utilized for interpreting and explaining model outputs by attributing a SHAP value to each feature, which indicates the feature's contribution to the final prediction in comparison to a baseline.

In this paper, we present a novel perspective on SHAP by integrating it as a new FW approach in clustering. Instead of relying on model-specific interpretations, we leverage the core concept behind SHAP values, quantifying each feature's contribution to assign data-driven weights that emphasize feature relevance within an unsupervised context. This adaptation of SHAP diverges from its traditional supervised applications and is fundamentally different from existing FW methods that utilize internal clustering metrics or heuristics.

In addition to using stand-alone SHAP values as weights, we combine them as an ensemble with other known FW methods. These combined approaches can surpass the performance of individual methods and even improve the overall effectiveness of unsupervised clustering algorithms.

2. Related Work

The use of FW in clustering has been studied extensively. Early work primarily focused on modifying existing clustering algorithms to assign and update feature weights during the clustering process. For instance, in weighted k-means, each feature is given a weight that is adapted iteratively to minimize within-cluster variance, aiming to place higher emphasis on features that are more discriminative [8].

Similarly, Ward's method [4], originally introduced for hierarchical clustering, has been extended to account for feature-specific weights (sometimes referred to as Ward variants) by adjusting the distance measure used in building the hierarchy, such as the Minkowski distance [9].

Other works have tackled FW by separating it from the clustering procedure itself (i.e., using filter approaches). Filter-based methods rely on statistical tests or correlations to rank features based on their intrinsic properties, which can be representative of potential cluster structures [5]. In [10], a method for feature weighting called "K-means Clustering-based Feature Weighting" was suggested. This method first extracts features from the frequency domain and calculates their mean, minimum, maximum, and standard deviation as statistical measurements. In the next stage, the K-means algorithm groups these features together, and the average values of the features in relation to the centers of these groups are used

as weights. Gürüler et al. [11] propose a hybrid system that integrates a complex-valued artificial neural network with a feature weighting technique based on k-means clustering. The clustering method is used to assign importance to input features by analyzing their distribution and separability across clusters, which improves the network's ability to discriminate between Parkinson's and non-Parkinson's cases. This method also demonstrates how unsupervised clustering can be effectively combined with neural network models to enhance feature relevance and classification performance.

While these filter FW approaches can be computationally efficient and widely applicable, they do not take into account the behavior of a specific clustering algorithm. In [5], an extensive classification of FW research works is presented; it is stated that the global filter FW approach in unsupervised learning is not a commonly employed method and that few works are encountered in the literature.

In recent years, XAI techniques such as SHAP have transformed how researchers interpret model decisions [6,7]. SHAP decomposes predictions into additive feature contributions, making it possible to explain complex models in a manner consistent with game-theoretic axioms. Although SHAP has primarily been used in supervised learning settings (classification and regression), its underlying principle—quantifying each feature's marginal contribution—is promising for providing FW in unsupervised learning. A few studies have started to explore combining XAI methods with cluster analysis, mostly for model interpretability or cluster labeling [12,13]. However, leveraging SHAP directly to derive feature weights that improve clustering outcomes remains largely unexplored territory.

Our work bridges this gap by introducing a SHAP-based global filter FW approach specifically tailored for clustering. By integrating the core ideas of SHAP values into the feature selection and weighting process, we aim to produce meaningful weights that not only enhance clustering performance but also provide what SHAP was meant to offer initially, i.e., feature importance.

3. Materials and Methods

Our primary objective is to integrate the numerical values derived from SHAP into the FW process for clustering tasks. SHAP typically provides a measure of each feature's marginal contribution to a predictive model's output in a supervised context. We adapt this concept to unsupervised tasks by training a surrogate predictive model (e.g., a classification model derived from the pseudo-labels) on an initial prediction Y_0, made by a clustering algorithm.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the data matrix, and let a clustering algorithm $\mathcal{C}$ produce an initial prediction Y_0. To temper dependence on the quality of Y_0, we perform multiple random restarts and select a well-initialized Y_0 in a single-pass scheme. In preliminary runs, weight rankings were stable across restarts, and overall performance varied only within baseline variability. For clarity and reproducibility, we therefore proceed with one representative Y_0.

We then train a random forest classifier f [14] to predict Y_0 from $\mathbf{X}$. On tabular data, forests capture non-linearities and interactions without extensive tuning and enable the use of TreeSHAP. We compute SHAP values with `TreeExplainer` [7], which yields exact, polynomial-time Shapley values for tree ensembles and avoids the variance and background-set sensitivity of model-agnostic approximations.

TreeSHAP returns per-sample, per-class attributions $\phi_{i,k}(x_j; f)$ for feature $i \in \{1, \ldots, d\}$, class $k \in \{1, \ldots, K\}$, and sample $j \in \{1, \ldots, n\}$. We aggregate to a single score per feature via the mean absolute SHAP [6]:

$$s_i = \frac{1}{nK} \sum_{j=1}^{n} \sum_{k=1}^{K} \left| \phi_{i,k}(x_j; f) \right|, \qquad w_i^{\text{SHAP}} = \frac{s_i}{\sum_{p=1}^{d} s_p}. \tag{1}$$

Taking absolute values focuses on magnitude rather than direction and is a common and empirically supported practice for feature importance with SHAP [7]. The resulting normalized vector $\mathbf{W}$ is then used to rescale features:

$$\widetilde{\mathbf{X}} = \mathbf{X} \cdot \mathbf{W}. \tag{2}$$

We reapply $\mathcal{C}$ to $\widetilde{\mathbf{X}}$ to obtain predictions Y. The entire process is summarized visually in Figure 1. Although the weighting can be iterated in a wrapper-like fashion by recomputing SHAP on Y until a stopping rule is met, in our experiments, we use a single iteration to isolate the causal effect of SHAP-derived weights, preserve comparability across algorithms, and maintain computational efficiency. This also avoids potential bias from repeatedly evaluating on the same pseudo-labels.

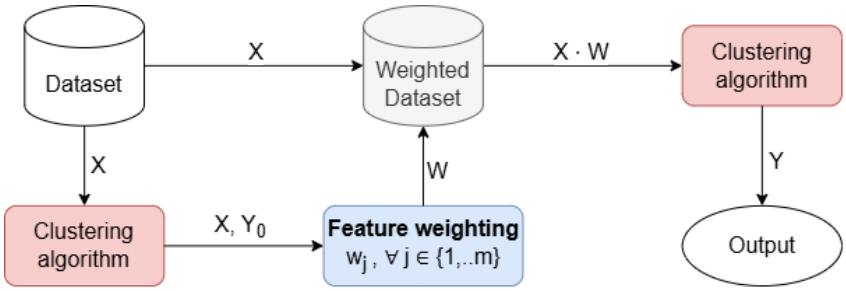

Figure 1. Flowchart of the employed FW methodology.

The resulting SHAP-based weights can then be combined with traditional FW strategies in an ensemble by multiplying the weights and applying them to the data. Multiplication amplifies consensus (features valued by both strategies) and suppresses disagreement (features valued by only one), which makes the metric space sharper along globally discriminative axes. This ensemble weighting strategy aims to retain each method's strengths and even surpass the performance of the stand-alone method.

Let $\mathbf{W}^{(g)} \in \mathbb{R}^d_+$ be normalized weights from another FW strategy g, $\sum_i w_i^{(g)} = 1$. We combine SHAP and g by elementwise multiplication followed by renormalization:

$$w_i^{\text{ens}} = \frac{w_i^{(\text{SHAP})} w_i^{(g)}}{\sum_{p=1}^{d} w_p^{(\text{SHAP})} w_p^{(g)}}, \qquad i = 1, \ldots, d. \tag{3}$$

Multiplication amplifies consensus (features valued by both methods) and suppresses disagreement (features emphasized by only one), sharpening the metric space along globally discriminative axes.

We integrate and evaluate SHAP as an FW method alongside several other FW or feature selection methods adapted as FW, ensuring that the selected approaches are most varied in terms of their underlying principles, methodologies, and mathematical foundations:

1. Minkowski distance (L_p norm). Inspired by [9], this approach can be considered a generalization of both the Euclidean distance ($p = 2$) and the Manhattan distance ($p = 1$) between two points.
2. Minimum Redundancy Maximum Relevance (mRMR). A feature selection method adapted for FW, aiming to maximize the relevance of selected features to the target variable (or pseudo-label in unsupervised cases) while minimizing redundancy among features [15].
3. Principal Component Analysis (PCA). A technique also used mainly in feature selection and dimensionality reduction. We adapt the principal component loadings as

a proxy for feature importance. Larger loadings suggest a stronger influence on the principal components [16].

4. One-way analysis of variance (F-test statistic). Often used as a method to compare statistical models, it can be adapted to act as an FW method. It is represented by the ratio of two scaled sums of squares reflecting different sources of variability [17].
5. *t*-Distributed Stochastic Neighbor Embedding (t-SNE). Another unsupervised non-linear dimensionality reduction technique, embedding high-dimensional points in low dimensions in a way that respects similarities between points [18].

In order to run experiments for evaluating the performance of the FW strategies, we employed a hosted T4 GPU provided by Google Colab. For acquiring datasets and implementing algorithms, as well as for using common evaluation metrics, we used the open-source library scikit-learn [19], along with datasets from the UCI repository [20]. All clustering algorithms were used with their default parameters as specified in scikit-learn.

3.1. Datasets

We conduct experiments on well-known datasets in the field of machine learning, which can be adapted easily for clustering by ignoring the target feature during the clustering process:

1. Diabetes (Pima) dataset: Contains 768 samples across 8 features that describe patients at risk of diabetes, with the objective of predicting the presence of the disease [21].
2. Wine recognition dataset: Consists of 178 samples characterized by 13 chemical analysis features of wines derived from three different cultivars, resulting in three classes [22].
3. Breast Cancer Wisconsin (diagnostic) dataset: Features 569 samples with 30 features describing tumor cells from clinical samples labeled as benign or malignant, resulting in 2 classes [23].
4. Optical recognition of handwritten digits dataset: Contains 1797 images of handwritten digits, resulting in 10 classes where each class refers to a digit [24].
5. Vehicle Silhouettes: Contains 946 instances for classifying a given vehicle as one of four types, using a set of 18 features extracted from their silhouette [25].

Although class labels exist in the datasets, we utilize them only at the evaluation stage to compute external clustering metrics. The clustering itself remains unsupervised.

3.2. Clustering Algorithms

Four different common clustering algorithms are employed:

1. *k*-means, a centroid-based algorithm that partitions data into *k* clusters by minimizing within-cluster variance [1];
2. Hierarchical clustering (Ward's method), a bottom-up approach that successively merges clusters to minimize the increase in sum-of-squares [4];
3. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), a density-based clustering algorithm that can handle varying densities, forming a hierarchical tree of possible clusters and extracting stable subclusters [26];
4. Gaussian Mixture Models (GMM), a model-based technique assuming data are generated from a finite mixture of Gaussians, optimized via the Expectation–Maximization (EM) algorithm [2].

3.3. Evaluation Metrics

We evaluate clustering performance using four metrics that provide a comprehensive and balanced evaluation from different perspectives, both externally and internally:

1. Adjusted Rand Index (ARI) measures the similarity between the predicted clusters and ground-truth labels, adjusting for chance [27];
2. Silhouette Score quantifies how well samples in the same cluster are similar to each other compared to samples from other clusters [28];
3. Normalized Mutual Information (NMI) evaluates the amount of mutual information between cluster assignments and ground-truth classes, normalized to the range $[0, 1]$ [5];
4. Calinski–Harabasz Index (CH), also called the Variance Ratio Criterion, assesses the ratio of between-cluster dispersion to within-cluster dispersion [29].

While both ARI and NMI are external indices, ARI penalizes splits and merges more rigorously, whereas NMI is more tolerant of proportional splits. In contrast, Silhouette and CH are internal indices that evaluate the geometric properties of clusters. The Silhouette score emphasizes compact and well-separated clusters on a local level, while CH favors clusters with high between-cluster to within-cluster variance ratios, often preferring larger, more spherical, and balanced cluster structures.

Thus, a solution may achieve a high score on Silhouette or CH yet score low on ARI or NMI if it is geometrically well-organized but semantically misaligned. ARI can decrease significantly under over-segmentation, while NMI may remain moderate. Additionally, Silhouette and CH may yield differing evaluations due to factors such as cluster shape, outliers, class imbalance, and distance scaling. In practical terms, the results can be summarized as a trade-off between semantic fidelity and geometric separability.

4. Results

To present our results, we plotted the notable situations in Figures 2–4 where SHAP-based FW exceeds other FW methods. All results are presented in Appendices A.1–A.5, with each notable result related to SHAP highlighted. The lines of the result tables where multiple FW methods are enumerated denote a FW ensemble (multiplication).

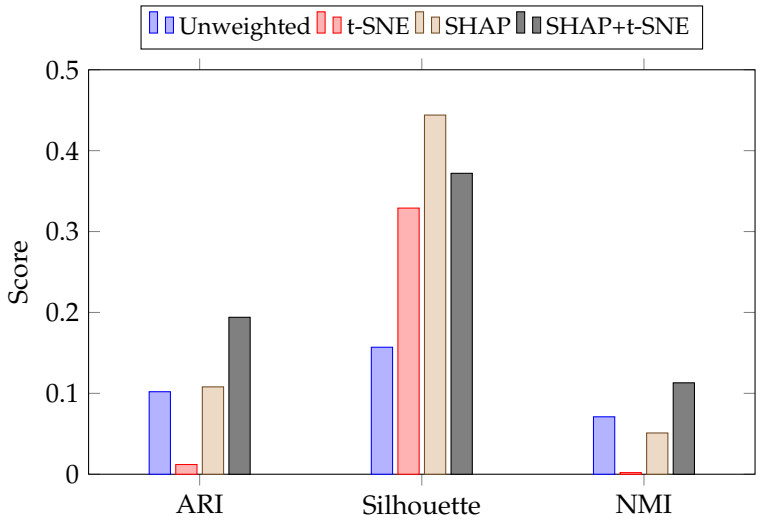

Figure 2. Hierarchical clustering metrics across three FW methods on the Diabetes dataset.

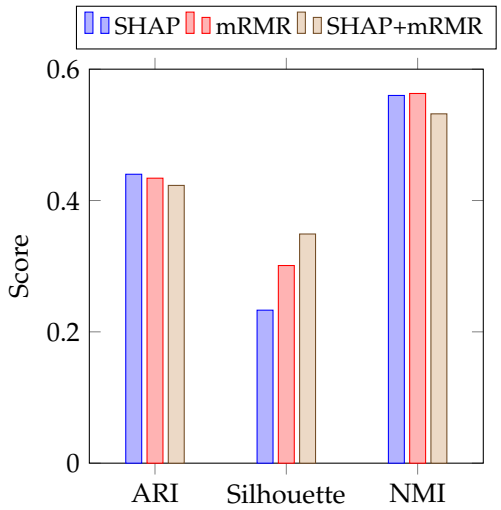

Figure 3. HDBSCAN metrics across three FW methods on the Wine dataset.

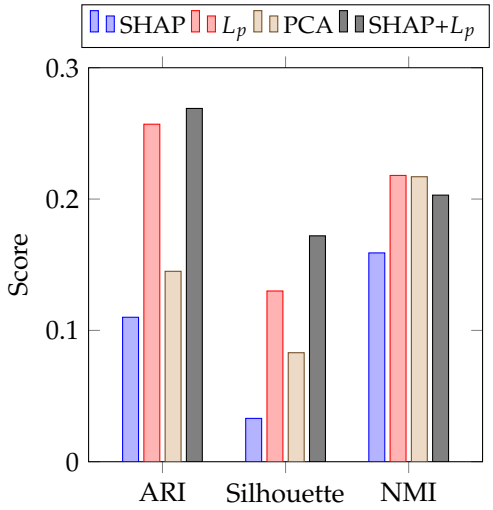

Figure 4. HDBSCAN metrics on four FW methods for the Breast Cancer dataset.

4.1. Diabetes (Pima) Dataset

For the diabetes dataset, as seen in Appendix A.1, the ensemble methods, specifically SHAP paired with t-SNE and with F-test, consistently offer gains in internal cluster quality (Silhouette, CH) and, in some cases, external metrics (ARI, NMI).

In the case of the *k*-means algorithm, a noteworthy observation is that the unweighted dataset presents an ARI of 0.101, which, while relatively low considering the definition of the evaluation metric, exceeds the performance of any of the FW approaches.

For hierarchical clustering, SHAP alone increases the ARI from 0.102 (unweighted) to 0.108 and greatly improves the Silhouette score (from 0.157 to 0.444). However, the ensemble combining SHAP with t-SNE jumps to an ARI of 0.194 and shows the highest NMI (0.113) among the FW approaches, indicating that pairing SHAP with dimensionality reduction can better isolate the underlying cluster structure. Other ensembles also improve over the baseline, but SHAP combined with t-SNE seems to be the most effective in terms of external metrics.

The ARI and NMI values of HDBSCAN remain quite low. In contrast, internal metrics like the Silhouette and especially the CH index show dramatic improvements, especially in the case of F-test and its SHAP ensemble. This discrepancy suggests that although some methods form very distinct clusters (as per the internal measures), these clusters do not align as well with the true class labels.

SHAP-based weighting alone does not uniformly improve the ARI, especially in methods like *k*-means and GMM, where unweighted or other feature selection techniques (F-test) sometimes perform better.

Overall, these findings indicate that while SHAP delivers satisfactory performance, its combination with other strategies can further enhance the quality of clustering.

4.2. Wine Dataset

SHAP-based weighting consistently performs well across *k*-means and GMM. Hierarchical clustering sees a stark improvement with the L_p method alone, as shown in Appendix A.2.

Regarding *k*-means, the ensembles involving SHAP do not strongly surpass the single SHAP approach; in fact, the ensembles have slightly lower ARIs compared to SHAP alone. This might indicate that SHAP's weighting is already well aligned with the relevant wine features.

When it comes to hierarchical clustering, SHAP alone has an ARI of 0.601, while SHAP combined with PCA jumps to 0.832 in ARI and 0.820 in NMI. This improvement suggests that combining SHAP weights with PCA loadings can better discriminate hierarchical clusters.

In the context of HDBSCAN, mRMR stands out, with a high ARI (0.434) and the best Silhouette among single methods (0.301), and SHAP alone performs similarly (0.440 ARI, 0.233 Silhouette), as seen in Figure 3. The ensemble methods do not yield substantial improvements here in ARI, but do improve Silhouette.

When applying GMM, SHAP has an ARI of 0.947 and an NMI of 0.928, which is near the top. The highest ARI across all strategies is from SHAP alone (0.947), while the ensembles of SHAP–L_p and SHAP–PCA remain close but slightly lower in ARI.

4.3. Breast cancer Wisconsin Dataset

As seen in Appendix A.3, SHAP and mRMR appear to be the two most reliable strategies across *k*-means, hierarchical, and GMM. Both of them combined sometimes help (particularly with hierarchical clustering) but can harm GMM performance. F-test systematically shows high internal metrics (Silhouette, CH) but fails to align well with external metrics.

When applying *k*-means, SHAP performs reasonably well (0.659 ARI, 0.548 NMI). Notably, the SHAP–mRMR ensemble yields a high ARI of 0.671 and a high Silhouette score of 0.588, suggesting a beneficial combination.

Regarding hierarchical clustering, SHAP performs the best out of the tested methods in terms of ARI (0.719) and NMI (0.599). Moreover, SHAP combined with mRMR stands out with a high ARI of 0.694 and the highest CH (1029.995), indicative of well-separated clusters.

In the context of HDBSCAN, it is notable that the SHAP–L_p ensemble performs the best in terms of ARI (0.269) and CH (72.904), as observed in Figure 4. Another ensemble, SHAP–F-test, provides the highest Silhouette score (0.353).

Applying GMM, SHAP leads with both ARI (0.793) and NMI (0.682). Just like in the previous algorithm's case, the SHAP–F-test ensemble has the highest Silhouette score, 0.634.

4.4. Digits Dataset

As observed in Appendix A.4, mRMR tends to dominate *k*-means and HDBSCAN with higher ARI, while SHAP excels for hierarchical clustering.

For other datasets, F-test yields extremely high internal metrics (Silhouette and CH) but very low external agreement (ARI and NMI), suggesting that internal compactness can be misleading when the underlying label distribution is complex.

As this dataset contains 10 clusters, we opted to omit t-SNE, as the number of components should be fewer than 4 for the Barnes–Hut underlying algorithm of t-SNE to function efficiently.

While combining SHAP weights with other filter/feature selection techniques yielded synergy on some datasets (e.g., Wine), it did not consistently improve the digit clustering. Instead, single-method approaches (SHAP or mRMR) frequently performed better, indicating that synergy benefits highly depend on the data distribution.

4.5. Vehicle Silhouette Dataset

For this dataset, t-SNE was excluded again due to the high number of clusters. The results in Appendix A.5 demonstrate that SHAP markedly improves clustering performance across most of the clustering algorithms. Compared to the unweighted dataset, SHAP consistently boosts the ARI and the NMI.

For example, in the case of k-means, the unweighted ARI is 0.075, while using SHAP increases it to 0.096. Similar trends are observed for other metrics.

The SHAP and L_p approach, in particular, yields the best overall performance, with an ARI of 0.122 and an NMI of 0.168 for k-means, an ARI of 0.133 and an NMI of 0.214 for hierarchical clustering, and 0.168 ARI and 0.334 NMI for the GMM algorithm.

This synergy further emphasizes SHAP's ability to capture nuanced, potentially nonlinear feature importance information, which complements the strengths of the L_p norm.

5. Discussion

In general, SHAP alone often delivers competitive ARI and NMI values across all used datasets, and in some cases surpasses other methods. The synergy with other methods (most notably SHAP–mRMR and SHAP–L_p) can improve certain algorithms, but it can also slightly degrade the stand-alone FW methods.

Empirical patterns show that SHAP and its ensemble counterparts perform well when the dataset exhibits globally informative structure. For example, in the Wine and Breast Cancer datasets, SHAP-based ensembles significantly improved both internal and external clustering metrics. These datasets are characterized by a small number of dominant axes that are useful for separating all classes. In such settings, the multiplicative fusion sharpens separation along relevant dimensions without excluding important complementary cues, aligning geometric separability with semantic consistency.

However, not all datasets benefit equally. In the Diabetes and Vehicle datasets, performance gains were limited or inconsistent. These datasets tend to have a more heterogeneous structure, where different clusters rely on distinct subsets of features. In such cases, multiplying SHAP with another score can inadvertently suppress features critical for specific clusters, especially if those features are not globally ranked highly by both sources. As a result, internal metrics such as Silhouette or CH may improve due to better compactness and separation, while external indices like ARI or NMI may stagnate or decline due to a misalignment with ground-truth labels.

Another recurring phenomenon is the emergence of inflated internal scores alongside degraded semantic alignment in high-dimensional or imbalanced datasets. This is particularly evident in the Digits dataset, where certain combinations of weighting strategies (e.g., SHAP with F-test) led to near-perfect internal metrics but low external agreement. These results reflect overconcentration of weights on a small number of features, distorting distance computations and inducing artificial separation that is not semantically meaningful.

The observed variability in ensemble performance also suggests that the geometric properties of the data interact with the weighting scheme. For example, datasets with high feature redundancy or collinearity may suffer when SHAP is combined with redundancy-penalizing methods like mRMR. While mRMR is designed to suppress collinear features, doing so in conjunction with SHAP can lead to excessive pruning of informative dimensions, particularly in datasets with curved or manifold structure.

In terms of what algorithms are susceptible to improvement by SHAP alone, the situation is dataset-dependent:

- Diabetes dataset: Hierarchical clustering, HDBSCAN, and GMM;
- Wine dataset: *k*-means, HDBSCAN, and GMM;
- Breast Cancer dataset: Hierarchical clustering and GMM;
- Digits dataset: Hierarchical clustering;
- Vehicle Silhouette dataset: *k*-means, hierarchical clustering, and GMM.

As a practical implication, our findings confirm that no single weighting strategy universally dominates. Rather, performance depends on the synergy between the dataset characteristics, the clustering algorithm, and the metric used. With this in mind, it can be observed that SHAP does not underperform, nor does it perform best under one single configuration. Therefore, practitioners can leverage SHAP as a general-purpose, reliable, and versatile FW method while simultaneously gaining insights into each feature's contribution.

6. Conclusions

In this paper, we presented feature weighting approaches for clustering, motivated by the need to identify and weight the most informative features during unsupervised learning in a new way. By adapting SHAP—originally designed for supervised settings—we leveraged SHAP values as a principled way to estimate each feature's contribution in distinguishing pseudo-clusters. Our proposed method was systematically compared against other FW strategies, namely L_p, mRMR, PCA, F-test, and t-SNE, both as stand-alone techniques and in combination with SHAP through ensemble weighting.

Experimental results on standard datasets (Diabetes, Wine, Breast Cancer, Digits, and Vehicle Silhouette) and four clustering algorithms (*k*-means, hierarchical clustering, HDBSCAN, and GMM) demonstrated that SHAP-based feature weighting frequently provides competitive performance, often approaching or outperforming established methods with respect to external clustering metrics like ARI and NMI, especially for data suited for binary clustering, like the Breast Cancer dataset. Moreover, in certain scenarios—especially for density-based or hierarchical approaches—combining SHAP with other methods (e.g., SHAP–mRMR or SHAP–L_p) proved beneficial in improving cluster separability, as reflected by internal metrics like Silhouette and CH in relatively well-separated clusters, for example, the Wine dataset. Nonetheless, we observed that these benefits are still dataset- and algorithm-dependent, but perform well enough for this approach to be considered general-purpose and reliable.

Despite promising results, limitations exist. First, deriving SHAP values for clustering involves building a pseudo-supervised setup on unlabeled data (training a model on generated labels), which can increase computational overhead for large datasets. Furthermore, SHAP-based weighting relies on how accurately pseudo-labels approximate the underlying cluster structure. If the surrogate model poorly reflects the natural groupings or if the pseudo-labeling process is unstable, the resulting weights may not be optimal. Additionally, while our experiments included multiple well-known datasets, testing on other domains or signal processing data [30,31] could further validate robustness and reveal additional edge cases. By addressing these directions, we aim to strengthen the theoretical foundations of

SHAP-inspired feature weighting in unsupervised learning and improve its utility in this way, not only as a tool purely for gaining explainability.

Author Contributions: Conceptualization, F.G., D.M.O. and C.I.; methodology, F.G., D.M.O. and C.I.; software, F.G., D.M.O. and C.I.; validation, F.G., D.M.O. and C.I.; formal analysis, F.G., D.M.O. and C.I.; investigation, F.G., D.M.O. and C.I.; resources, F.G., D.M.O. and C.I.; data curation, F.G., D.M.O. and C.I.; writing—original draft preparation, F.G., D.M.O. and C.I.; writing—review and editing, F.G., D.M.O. and C.I.; visualization, F.G., D.M.O. and C.I.; supervision, F.G., D.M.O. and C.I.; project administration, F.G., D.M.O. and C.I.; funding acquisition, F.G., D.M.O. and C.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by the project "Romanian Hub for Artificial Intelligence-HRIA", Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARI	Adjusted Rand Index
CH	Calinski–Harabasz (Index)
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FW	Feature Weighting
GMM	Gaussian Mixture Model
HDBSCAN	Hierarchical DBSCAN
L_p	L_p norm (Minkowski metric)
mRMR	Minimum Redundancy Maximum Relevance
NMI	Normalized Mutual Information
PCA	Principal Component Analysis
SHAP	SHapley Additive exPlanations
Sil	Silhouette score
t-SNE	*t*-distributed Stochastic Neighbor Embedding
XAI	eXplainable Artificial Intelligence

Appendix A. Complete Clustering Results

Appendix A.1. Diabetes (Pima) Dataset

	k-Means				**Hierarchical Clustering (Ward)**				**HDBSCAN**				**Gaussian Mixture Model**			
	ARI	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**
Unweighted	0.101	0.165	0.059	146.743	0.102	0.157	0.071	117.826	0.053	0.256	0.023	35.951	0.013	0.165	0.002	115.419
SHAP	0.001	0.368	0.009	538.785	0.108	0.444	0.051	619.946	0.021	0.494	0.013	145.707	0.013	0.600	0.002	1343.91
L_p	0.014	0.232	0.017	260.910	0.100	0.275	0.048	250.810	0.024	0.439	0.012	101.140	0.013	0.396	0.002	574.569
mRMR	0.001	0.501	0.006	1152.98	0.066	0.352	0.042	499.023	0.012	0.733	0.009	421.034	0.013	0.649	0.002	1704.41
PCA	0.096	0.187	0.052	170.693	0.139	0.229	0.078	133.408	0.046	0.256	0.019	34.111	0.013	0.174	0.002	128.979
F-test	0.089	0.665	0.041	2002.12	0.089	0.663	0.046	1955.205	0.017	0.961	0.042	6516.82	0.098	0.592	0.070	1375.37
t-SNE	0.002	0.302	0.008	340.947	0.012	0.329	0.002	311.014	0.015	0.436	0.005	76.060	0.013	0.329	0.002	312.241
SHAP, L_p	0.003	0.441	0.006	773.292	0.078	0.466	0.045	868.172	0.034	0.355	0.017	226.833	0.013	0.654	0.002	1681.47
SHAP, mRMR	0.005	0.589	0.001	1591.97	0.110	0.523	0.057	1054.11	0.028	0.440	0.013	324.799	0.013	**0.673**	0.002	**1754.12**
SHAP, PCA	0.002	0.364	0.009	517.279	0.089	0.423	0.039	558.413	0.021	0.487	0.013	147.734	0.013	0.595	0.002	1282.64
SHAP, F-test	0.089	0.665	0.041	2002.12	0.089	0.663	0.046	1955.21	0.018	**0.963**	**0.043**	**8498.43**	0.098	0.592	0.070	1375.37
SHAP, t-SNE	0.005	0.513	0.001	1137.39	**0.194**	0.372	**0.113**	377.202	0.012	0.726	0.003	726.664	0.013	0.645	0.002	1653.94

Note: Bold values indicate notable results involving SHAP (as a stand-alone method or in an ensemble).

Appendix A.2. Wine Dataset

	k-Means				Hierarchical Clustering (Ward)				HDBSCAN				Gaussian Mixture Model			
	ARI	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**
Unweighted	0.871	0.284	0.875	70.940	0.789	0.277	0.786	67.647	0.345	0.133	0.449	24.689	0.897	0.284	0.875	70.940
SHAP	**0.880**	0.435	0.850	160.129	0.601	0.364	0.654	136.720	**0.440**	0.233	0.560	62.690	**0.947**	0.429	**0.928**	155.921
L_p	0.790	0.370	0.770	139.341	0.964	0.351	0.954	137.720	0.395	0.262	0.485	57.968	0.852	0.365	0.836	135.390
mRMR	0.820	0.396	0.799	153.493	0.817	0.394	0.794	165.233	0.434	0.301	0.563	85.362	0.915	0.386	0.893	146.743
PCA	0.881	0.283	0.865	69.566	0.730	0.265	0.716	64.628	0.341	0.130	0.425	24.450	0.897	0.275	0.876	67.308
F-test	0.374	0.591	0.453	487.088	0.320	0.603	0.392	462.742	0.122	0.389	0.269	25.868	0.318	0.599	0.394	470.792
t-SNE	0.698	0.322	0.706	126.626	0.837	0.312	0.815	117.914	0.390	0.231	0.493	54.217	0.756	0.316	0.762	122.395
SHAP, L_p	0.804	0.450	0.784	195.976	0.588	0.415	0.652	251.179	0.423	0.274	0.539	87.190	0.915	0.441	0.893	186.088
SHAP, mRMR	0.834	0.446	0.807	183.181	0.712	0.408	0.743	208.013	0.423	0.349	0.532	**111.704**	0.897	0.430	0.876	171.891
SHAP, PCA	0.880	0.433	0.850	157.875	0.832	0.408	0.820	150.307	0.403	0.206	0.507	50.067	0.931	0.428	0.909	153.549
SHAP, F-test	0.374	0.591	0.453	487.088	0.320	0.603	0.392	462.742	0.104	**0.398**	0.256	18.604	0.318	0.599	0.394	470.792
SHAP, t-SNE	0.847	0.424	0.815	187.333	0.672	0.411	0.710	235.881	0.392	0.226	0.534	65.373	0.895	0.422	0.882	185.763

Note: Bold values indicate notable results involving SHAP (as a stand-alone method or in an ensemble).

Appendix A.3. Breast Cancer Dataset

	k-Means				Hierarchical Clustering (Ward)				HDBSCAN				Gaussian Mixture Model			
	ARI	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**
Unweighted	0.676	0.344	0.562	267.680	0.575	0.339	0.456	248.628	0.156	0.028	0.212	48.146	0.774	0.314	0.661	247.283
SHAP	0.659	0.575	0.548	944.123	**0.719**	0.542	0.599	880.312	0.110	0.033	0.159	**65.027**	**0.793**	0.521	**0.682**	752.406
L_p	0.718	0.449	0.617	495.080	0.586	0.419	0.464	411.237	0.257	0.130	0.218	56.893	0.755	0.419	0.641	447.810
mRMR	0.730	0.487	0.629	617.511	0.707	0.453	0.585	554.375	0.000	0.000	0.000	0.000	0.767	0.461	0.655	570.654
PCA	0.642	0.350	0.523	269.743	0.603	0.339	0.479	258.376	0.145	0.083	0.217	33.686	0.660	0.332	0.537	247.117
F-test	0.126	0.618	0.070	920.864	0.127	0.611	0.071	908.811	0.001	0.352	0.046	27.081	0.137	0.633	0.082	921.494
SHAP, L_p	0.653	0.592	0.542	1068.392	0.466	0.574	0.421	900.122	**0.269**	0.172	0.203	**72.904**	0.118	0.484	0.115	248.153
SHAP, mRMR	0.671	0.588	0.559	**1069.674**	0.694	0.577	0.592	**1029.995**	0.000	0.000	0.000	0.000	0.174	0.456	0.159	313.117
SHAP, PCA	0.659	0.573	0.548	924.684	0.689	0.546	0.568	897.500	0.076	0.054	0.145	57.197	0.103	0.513	0.114	224.856
SHAP, F-test	0.126	0.618	0.070	920.864	0.127	0.611	0.071	908.811	0.001	**0.353**	0.047	21.695	0.139	**0.634**	0.084	918.215
SHAP + t-SNE	0.688	0.581	0.583	1004.840	0.412	0.552	0.372	766.806	0.361	0.172	0.283	88.030	0.370	0.528	0.333	590.184

Note: Bold values indicate notable results involving SHAP (as a stand-alone method or in an ensemble).

Appendix A.4. Digits Dataset

	k-Means				Hierarchical Clustering (Ward)				HDBSCAN				Gaussian Mixture Model			
	ARI	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**
Unweighted	0.530	0.135	0.672	113.060	0.664	0.125	0.795	105.825	0.209	0.041	0.580	30.762	0.546	0.117	0.690	102.577
SHAP	0.502	0.210	0.626	243.155	**0.700**	0.181	0.786	215.512	0.376	0.005	0.650	69.313	0.518	0.170	0.651	241.091
L_p	0.402	0.103	0.546	5.82×10^{16}	0.531	0.154	0.738	184.417	0.001	0.334	0.048	30.499	0.000	1.000	0.001	3.53×10^{17}
mRMR	0.560	0.192	0.660	201.636	0.667	0.185	0.788	184.384	0.454	0.083	0.724	77.951	0.615	0.199	0.717	196.626
PCA	0.511	0.134	0.660	115.868	0.660	0.136	0.778	112.902	0.219	0.032	0.584	33.854	0.557	0.128	0.687	113.244
F-test	0.003	0.986	0.068	49,688	0.003	0.984	0.068	69,182	0.003	0.992	0.070	131,083	0.003	0.986	0.068	50,981
SHAP, L_p	0.441	0.199	0.566	1.38×10^{12}	0.635	0.185	0.730	237.468	0.001	0.976	0.042	232.357	0.000	0.000	0.000	0.000
SHAP, mRMR	0.487	0.220	0.595	274.584	0.547	0.187	0.669	256.476	0.337	0.016	0.668	73.446	0.505	0.211	0.628	302.152
SHAP, PCA	0.514	0.213	0.634	250.872	0.543	0.180	0.676	218.276	0.356	0.022	0.678	67.599	0.091	0.189	0.193	453.015
SHAP, F-test	0.003	0.986	0.068	49,688	0.003	0.984	0.068	60,631	0.003	**0.993**	0.070	**141,957**	0.003	0.986	0.068	49,688

Note: Bold values indicate notable results involving SHAP (as a stand-alone method or in an ensemble).

Appendix A.5. Vehicle Silhouette Dataset

	k-Means				Hierarchical Clustering (Ward)				HDBSCAN				Gaussian Mixture Model			
	ARI	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**	**ARI**	**Sil**	**NMI**	**CH**
Unweighted	0.075	0.299	0.121	390.829	0.055	0.265	0.104	353.824	0.002	0.585	0.022	66.224	0.084	0.280	0.130	368.677
SHAP	0.096	0.364	**0.141**	563.310	**0.128**	0.345	**0.172**	651.226	0.002	0.857	0.024	483.520	**0.141**	0.345	**0.204**	605.251
L_p	0.072	0.345	0.116	681.217	0.077	0.363	0.148	636.358	0.002	0.843	0.024	428.710	0.077	0.331	0.120	681.624
mRMR	0.114	0.298	0.130	575.250	0.126	0.313	0.165	604.981	0.002	0.758	0.024	210.294	0.104	0.180	0.182	394.014
PCA	0.076	0.327	0.122	437.053	0.110	0.316	0.148	413.983	0.002	0.597	0.022	69.256	0.085	0.298	0.131	400.975
F-test	0.051	0.576	0.057	4120.081	0.046	0.551	0.059	3821.233	0.018	0.997	0.070	709671	0.046	0.563	0.055	3952.806
SHAP, L_p	**0.122**	0.379	**0.168**	848.862	**0.133**	0.409	**0.214**	1080.091	0.002	0.909	0.024	1093.220	**0.168**	0.392	**0.334**	505.685
SHAP, mRMR	0.083	0.300	0.108	570.853	0.093	0.336	0.127	794.904	0.002	0.887	0.024	746.153	0.086	0.343	0.126	426.536
SHAP, PCA	0.083	0.360	0.106	582.401	0.101	0.378	0.172	598.437	0.002	0.853	0.024	466.502	0.082	0.392	0.123	451.038
SHAP, F-test	0.051	0.576	0.057	4120.081	0.039	0.540	0.056	3793.931	0.018	0.997	0.070	709671	0.046	0.563	0.055	3952.806

Note: Bold values indicate notable results involving SHAP (as a stand-alone method or in an ensemble).

References

1. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
2. Xu, R.; Wunsch, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [CrossRef] [PubMed]
3. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 2–4 August 1996; KDD'96; pp. 226–231.
4. Ward, J.H., Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
5. Niño-Adan, I.; Manjarres, D.; Landa-Torres, I.; Portillo, E. Feature weighting methods: A review. *Expert Syst. Appl.* **2021**, *184*, 115424. [CrossRef]
6. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
7. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
8. Modha, D.S.; Spangler, W.S. Feature weighting in k-means clustering. *Mach. Learn.* **2003**, *52*, 217–237. [CrossRef]
9. De Amorim, R.C. Feature relevance in ward's hierarchical clustering using the L p norm. *J. Classif.* **2015**, *32*, 46–62. [CrossRef]
10. Güneş, S.; Polat, K.; Yosunkaya, Ş. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Syst. Appl.* **2010**, *37*, 7922–7928. [CrossRef]
11. Gürüler, H. A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. *Neural Comput. Appl.* **2017**, *28*, 1657–1666. [CrossRef]
12. Yang, H.; Jiao, L.; Pan, Q. A survey on interpretable clustering. In Proceedings of the 2021 40th Chinese Control Conference (CCC), Shanghai, China, 26–28 July 2021; pp. 7384–7388.
13. Hu, L.; Jiang, M.; Dong, J.; Liu, X.; He, Z. Interpretable Clustering: A Survey. *arXiv* **2024**, arXiv:2409.00743. [CrossRef]
14. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
15. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
16. Jolliffe, I.T. Principal component analysis and factor analysis. In *Principal Component Analysis*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 150–166.
17. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
18. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
19. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
20. Asuncion, A.; Newman, D. UCI Machine Learning Repository. 2007. Available online: https://archive.ics.uci.edu/ (accessed on 12 August 2025)
21. Smith, J.W.; Everhart, J.E.; Dickson, W.C.; Knowler, W.C.; Johannes, R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care, Washington, DC, USA, 6–9 November 1988; p. 261.
22. Aeberhard, S.; Coomans, D.; De Vel, O. *Comparison of Classifiers in High Dimensional Settings;* Technical Report No. 92-02; Department of Mathematics and Statistics, James Cook University of North Queensland: North Queensland, Australia 1992.
23. Street, W.N.; Wolberg, W.H.; Mangasarian, O.L. Nuclear feature extraction for breast tumor diagnosis. In Proceedings of the Biomedical Image Processing and Biomedical Visualization, San Jose, CA, USA, 1–4 February 1993; Volume 1905, pp. 861–870.
24. Kaynak, C. Methods of Combining Multiple Classifiers and Their Applications to Handwritten Digit Recognition. Master Thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University, Istanbul, Turkey, 1995.
25. Siebert, J.P. Vehicle Recognition Using Rule Based Methods; Turing Institute: London, UK, 1987.
26. Campello, R.J.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Gold Coast, Australia, 14–17 April 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.
27. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
28. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
29. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [CrossRef]

30. Onchis, D. Observing damaged beams through their time–frequency extended signatures. *Signal Process.* **2014**, *96*, 16–20. [CrossRef]
31. Feichtinger, H.; Onchiş, D. Constructive reconstruction from irregular sampling in multi-window spline-type spaces. In *Progress in Analysis and Its Applications*; World Scientific: Singapore 2010; pp. 257–265.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.