AN IMPROVEMENT OF THE PICTURE-SAFE SEMI-SUPERVISED FUZZY CLUSTERING FOR PARTITIONING NOISY DATA

Pham Huy Thong[1,*], Phung The Huan[2]

DOI: https://doi.org/10.57001/huih5804.2025.417

ABSTRACT

This paper introduces an improved safe semi-supervised fuzzy clustering algorithm based on Picture Fuzzy Sets (PFS), named New_NPFS3FCM, designed to enhance clustering performance in noisy datasets. The proposed method introduces several key improvements, including selective updating of only the affected data points and clusters during label assignment, reusing precomputed distance values to eliminate redundant calculations, incorporating an early stopping criterion based on label stability, and excluding clusters without variation to focus computational resources on dynamic regions. Furthermore, the traditional Euclidean distance is replaced with a kernel-based distance in the feature space, enabling better modeling of complex nonlinear cluster structures. The algorithm was evaluated on 15 datasets from the UCI and ODDS repositories, covering both low-noise and high-noise scenarios, and compared against two established approaches, FC-PFS and CS3FCM. Experimental results demonstrate that New_NPFS3FCM achieves the highest classification accuracy in most datasets, with remarkable performance on labeled and noisy data, produces higher-quality clusters as reflected in lower Davies–Bouldin (DB) indices for 10 out of 15 datasets, and delivers competitive computational efficiency, particularly for medium-to-large datasets. Leveraging the expressive power of PFS in uncertainty modeling and the safety of semi-supervised integration, the proposed method demonstrates stability, noise resistance, and high effectiveness, making it well-suited for a wide range of real-world noisy data clustering tasks.

Keywords: *Fuzzy clustering, Safe semi-supervised fuzzy clustering, Kernel-based clustering, Picture fuzzy clustering, Data partition with noises.*

[1]Hanoi University of Industry, Vietnam

[2]Thai Nguyen University of Information and Communication Technology, Vietnam

[*]Email: thongph@haui.edu.vn

Received: 15/8/2025

Revised: 27/9/2025

Accepted: 28/11/2025

1. INTRODUCTION

Data clustering [1] is a core method in data mining and AI, used in customer segmentation, anomaly detection, image analysis, and bioinformatics. Its goal is to group objects so that members are more similar to each other than to those in other clusters. In practical settings, especially with data from natural environments or sensors, noise and uncertainty make clustering challenging and can degrade conventional algorithm performance. Classical approaches, such as K-means [2], often assume that every data point is equally reliable and contributes uniformly to the clustering process. This assumption can be misleading when the dataset contains outliers or erroneous entries, as it may distort the true cluster structure. Moreover, these methods typically employ "hard" partitioning, forcing each point into exactly one cluster, which is unsuitable for data with fuzzy or overlapping boundaries.

Fuzzy clustering offers a more adaptable alternative by allowing each data point to belong to multiple clusters with varying membership degrees. This flexibility makes it more effective when handling uncertainty or gradual transitions between clusters. Nevertheless, fully unsupervised fuzzy clustering can still produce unstable results on complex or noisy datasets, emphasizing the need to incorporate expert knowledge through semi-supervised learning. Semi-supervised learning (SSL) [3] uses a limited amount of labeled data to guide the clustering process, thereby improving accuracy and robustness. However, improper integration of labeled information may cause performance degradation compared to purely unsupervised approaches. This motivates the concept of "safe" semi-supervised learning, which ensures that labeled data is leveraged without harming clustering quality.

In this work, we introduce an improved safe semi-supervised fuzzy clustering algorithm based on Picture Fuzzy Sets (PFS) [4]. PFS extends traditional fuzzy sets by incorporating neutral and refusal degrees, enabling richer representation of uncertainty particularly

beneficial in scenarios with noise or ambiguous boundaries. By combining PFS with safe semi-supervised learning, the proposed method effectively harnesses expert knowledge while maintaining stability even in complex datasets. The proposed method presents three main advantages: (1) Robust modeling of noisy data through the expressive power of PFS; (2) Effective guidance from expert knowledge to improve clustering accuracy; and (3) Safe integration of labeled data without degrading performance compared to unsupervised methods. This combination is particularly well-suited for agricultural image analysis, where uncertainty and noise are common challenges.

2. RELATED WORK

Clustering has long been a cornerstone of research in computer science, particularly within the domains of machine learning, artificial intelligence, and image analysis [5]. Its primary objective is to reveal the latent organization of data without relying on prior label information, thereby enabling pattern discovery and exploratory analysis in situations where manual annotation is infeasible or prohibitively costly. The fundamental principle is straightforward: assemble data instances into internally coherent groups while ensuring that these groups remain as distinct as possible from one another. Yet, in the reality of practical applications, this principle is rarely easy to uphold. Real-world datasets are seldom clean; they often display overlapping categories, high variability in feature space, measurement noise, and boundary uncertainty. These conditions can severely impair the effectiveness of clustering algorithms that rely on simplistic partitioning rules.

Classical hard-assignment clustering methods, with K-means being the most widely recognized, have long been favored in both research and applied settings due to their notable advantages in computational efficiency, straightforward implementation, and intuitive interpretability [6]. These algorithms assign each data point exclusively to a single cluster, providing a crisp partitioning that is easy to analyze and visualize. Nevertheless, this very rigidity constitutes a fundamental limitation: they cannot accommodate partial or uncertain memberships, which are often inherent in real-world datasets. While hard-assignment approaches perform reasonably well when clusters are well-separated, roughly spherical, and homogenous, they struggle in more complex scenarios. In particular, when data categories overlap, exhibit irregular shapes, or when the boundaries between clusters are intrinsically ambiguous, these classical methods frequently fail to capture the nuanced structure of the data, highlighting the need for more flexible, uncertainty-aware clustering paradigms.

To address these limitations, the fuzzy clustering paradigm emerged, allowing data points to belong to multiple clusters with varying degrees of membership. This approach is particularly effective at capturing gradual transitions and soft boundaries, which are often observed in complex natural phenomena. Among fuzzy clustering techniques, Fuzzy C-Means (FCM) [7] has been the most widely adopted, in part because it strikes a favorable balance between conceptual simplicity and practical performance. Its optimization framework, based on minimizing within-cluster variance while respecting soft membership constraints, is flexible enough to adapt to a variety of data types. Nevertheless, standard FCM is not without its vulnerabilities: it can be sensitive to noise, is ill-equipped to model nonlinear cluster shapes, and is prone to distortion from outliers or irrelevant background information.

Over time, a range of FCM variants has been developed to overcome these weaknesses. Spatially-enhanced versions exploit the relationships between neighboring points, suppressing the influence of isolated noisy samples and mitigating lighting or texture variations in image data. Kernel-based variants perform an implicit nonlinear transformation of the data into a higher-dimensional feature space, enabling the algorithm to separate complex, non-convex cluster structures that are inseparable in the original space [8]. Robust FCM formulations incorporate alternative objective functions or membership models often inspired by statistical estimation theory or possibilistic logic that down-weight the contribution of anomalous points, thus improving stability in the presence of outliers.

Despite these advancements, clustering methods operating in a purely unsupervised manner are still bounded by the structural limitations inherent in the data. When class boundaries are unclear or when data noise overwhelms structural signals, the algorithm may converge to suboptimal partitions. SSL provides a compelling remedy by introducing a modest quantity of labeled data into the clustering process. This labeled subset, often provided by human experts, acts as a guiding signal to resolve ambiguities and steer the partitioning toward semantically meaningful clusters. SSL in clustering can be realized in multiple ways: through

constraint-based approaches that encode "must-link" and "cannot-link" relationships between sample pairs [9]; through metric learning techniques that adapt distance measures in accordance with labeled examples; or through graph-based label propagation schemes that leverage the connectivity of a similarity graph to extend known labels to unlabeled instances.

However, this integration of supervision is not without risks. If the labeled data is noisy, inconsistent, or incorrectly annotated, the clustering process can be misled, potentially producing worse results than a purely unsupervised approach. This problem has catalyzed the emergence of safe SSL frameworks methodologies designed so that the incorporation of labeled data never degrades the quality of the clustering and ideally leads to improvement when labels are reliable.

Parallel to the development of SSL, advancements in fuzzy set theory have introduced PFS, a generalization of conventional fuzzy sets that model three distinct membership degrees: positive, neutral, and negative. Unlike traditional fuzzy models that capture only the degree of belonging, PFS explicitly represents both acceptance and rejection, as well as an intermediate hesitancy degree. This richer representation is particularly beneficial when handling samples located near ambiguous or transitional regions of the feature space. The neutral membership captures indecision, allowing the clustering process to better accommodate borderline cases and complex overlaps that would otherwise be forced into binary membership decisions.

Despite the demonstrated advantages of PFS in handling uncertainty and noise, the integration of PFS with safe SSL remains largely unexplored. Most current work either applies PFS within purely unsupervised clustering frameworks or combines it with supervised cues without explicit safeguards against label noise [10]. This gap is especially significant for data environments where both advanced uncertainty modeling and cautious use of supervision are essential for example, datasets characterized by high heterogeneity, substantial measurement variability, and the presence of ambiguous samples.

From this perspective, three unresolved research challenges become apparent. First, existing clustering methodologies often fail to make full use of domain knowledge to disambiguate challenging cases. Second, SSL-enabled clustering approaches rarely implement robust safety mechanisms to prevent noisy supervision from eroding cluster quality. Third, while the theoretical compatibility between PFS and safe SSL is evident, systematic research on their integration is lacking. These factors collectively motivate the development of the Safe Semi-supervised Picture Fuzzy Clustering framework, a novel approach intended to merge the expressive capacity of PFS with the resilience of safe SSL. By doing so, the proposed method aims to deliver consistent performance in noisy conditions, handle ambiguous boundaries with greater precision, and integrate expert input without risking deterioration in overall clustering quality.

3. RESULTS AND DISCUSSION

3.1. Main Idea

The improved algorithm builds on the original method while introducing several key enhancements to boost efficiency and speed. Rather than processing the entire dataset in each iteration, it updates only the data points and clusters affected by changes during label assignment, significantly reducing computation time for large datasets. Previously computed distances are reused for points with unchanged positions relative to cluster centers, eliminating redundant calculations. A stability-based stopping criterion allows the algorithm to terminate early when labels remain constant across consecutive iterations. Additionally, clusters showing no internal variation are excluded from further processing, focusing resources on dynamic regions. These refinements not only accelerate execution but also enhance stability and accuracy, making the algorithm particularly effective for real-world tasks such as detecting and segmenting diseased plant leaves in large, noisy datasets.

3.2. Details of the NEW_NPFS3FCM

3.2.1. Proposed improvement

In the original NPFS3FCM algorithm (presented at the ICISN 2025 - International Conference on Intelligent Systems and Networks), the distance between a data point X_k and a cluster center V_j is computed using the Euclidean distance:

$$d_{kj}^2 = \left\| X_k - V_j \right\|^2 \quad (1)$$

When this is replaced by the distance in the kernel feature space ϕ, there is no need to compute ϕ instead, the kernel trick is applied:

$$\left\|\phi(X_k)-\phi(V_j)\right\|^2 = K(X_k,X_k) - 2\sum_{i=1}^{N} u_{ij}^m K(X_k,X_i)/\sum_{i=1}^{N} u_{ij}^m + \ldots \quad (2)$$

where $K(x,y)=\exp(-\frac{\|x-y\|^2}{2\delta^2}$ is a positive semi-definite kernel function, such as the Gaussian RBF kernel.

The original objective function is then improved by replacing the Euclidean distance term d_{kj}^2 with the kernel-based distance $d_{kj,K}^2$ defined in the kernel space:

$$d_{kj,K}^2 = K(X_k,X_k) - 2\frac{\sum_{i=1}^{N} u_{ij}^m K(X_k,X_i)}{\sum_{i=1}^{N} u_{ij}^m} + \frac{\sum_{i=1}^{N}\sum_{r=1}^{N} u_{ij}^m u_{rj}^m K(X_i,X_r)}{(\sum_{i=1}^{N} u_{ij}^m)^2} \quad (3)$$

This kernel-driven enhancement allows the algorithm to model intricate, non-linear cluster formations by performing an implicit transformation of the data into a higher-dimensional feature space, thereby avoiding the need for an explicit computation of the mapping ϕ.

Based on the aforementioned concepts, this section formalizes the proposed model through the following objective function:

$$F = \beta\sum_{k=1}^{N}\sum_{j=1}^{C}(\mu_{kj}(2-\xi_{kj}))^2 d_{kj,K}^2 - \delta\sum_{k=1}^{N}\sum_{j=1}^{C}\eta_{kj}(\ln(\eta_{kj})-\xi_{kj}) + \theta\sum_{k=1}^{L}\sum_{j=1}^{C}\frac{(\mu_{kj}(2-\xi_{kj})-f_{kj})^2}{1+(\bar{\mu}_{kj}-f_{kj})^2} d_{kj,K}^2 + \theta\sum_{k=L+1}^{N}\sum_{j=1}^{C}(\mu_{kj}(2-\xi_{kj})-\bar{\mu}_{kj})^2 d_{kj,K}^2 \to Min \quad (4)$$

With the following constraints:

$$\begin{cases} \mu_{kj},\xi_{kj},\eta_{kj}\in[0,1] \\ \mu_{kj}+\xi_{kj}+\eta_{kj}\le 1 \\ \sum_{j=1}^{C}(\mu_{kj})=1 \\ \sum_{j=1}^{C}\left(\eta_{kj}+\frac{\xi_{kj}}{C}\right)=1 \\ (k=\overline{1,N};\ j=\overline{1,C}) \end{cases} \quad (5)$$

Where data set $X=\{X_1,X_2,...,X_n\}$ with N elements, the number of labeled data in X: $L<N$; the cluster number: C; positive, neutral and refusal degrees of element X_k belong to cluster j: $\mu_{kj},\eta_{kj},\xi_{kj}$

Each component of the objective function serves a distinct role within the model's formulation. As illustrated in equation (1), the initial two terms correspond directly to those employed in the standard Picture Fuzzy Clustering (FC-PFS) [11] framework. In contrast, the remaining two terms encapsulate the safe semi-supervised clustering mechanism, specifically adapted to operate within the PFS representation.

Define the following variables:

$$A_{kj}^L = \left(\beta\left(2-\xi_{kj}\right)^2 d_{kj}^s + \frac{\theta\left(2-\xi_{kj}\right)^2 d_{kj}^s}{1+\left(\bar{\mu}_{kj}-f_{kj}\right)^2}\right) \quad (6)$$

$$B_{kj}^L = \sum_{i=1}^{C}\frac{A_{kj}^L}{A_{ki}^L} \quad (7)$$

$$G_{kj}^L = \frac{\theta f_{kj}\left(2-\xi_{kj}\right)d_{kj}^s}{1+\left(\bar{\mu}_{kj}-f_{kj}\right)^2} \quad (8)$$

$$A_{kj}^N = \left(\beta\left(2-\xi_{kj}\right)^2 d_{kj}^s + \theta\left(2-\xi_{kj}\right)^2 d_{kj}^s\right) \quad (9)$$

$$B_{kj}^N = \sum_{i=1}^{C}\frac{A_{kj}^N}{A_{ki}^N} \quad (10)$$

$$G_{kj}^N = \theta\bar{\mu}_{kj}\left(2-\xi_{kj}\right)d_{kj}^s \quad (11)$$

$$d_{kj}^S = X_k - V_j^2 \quad (12)$$

The optimal solutions, obtained by applying the Lagrangian method, are detailed in equations (10-14).

$$V_j = \frac{\beta\sum_{k=1}^{N}(\mu_{kj}(2-\xi_{kj}))^2 X_k + \theta\sum_{k=1}^{L}\frac{(\mu_{kj}(2-\xi_{kj})-f_{kj})^2}{1+(\bar{\mu}_{kj}-f_{kj})^2}X_k + \theta\sum_{k=L+1}^{N}(\mu_{kj}(2-\xi_{kj})-\bar{\mu}_{kj})^2 X_k}{\beta\sum_{k=1}^{N}(\mu_{kj}(2-\xi_{kj}))^2 X_k + \theta\sum_{k=1}^{L}\frac{(\mu_{kj}(2-\xi_{kj})-f_{kj})^2}{1+(\bar{\mu}_{kj}-f_{kj})^2} + \theta\sum_{k=L+1}^{N}(\mu_{kj}(2-\xi_{kj})-\bar{\mu}_{kj})^2} \quad (13)$$

The positive degree u of the labeled data elements is defined as follows:

$$\mu_{kj} = \frac{1}{B_{kj}^{L}} - \frac{\sum_{i=1}^{C} \frac{G_{ki}^{L}}{A_{ki}^{L}}}{B_{kj}^{L}} + \frac{G_{kj}^{L}}{A_{kj}^{L}} \quad (14)$$

The positive degree *u* of the unlabeled data elements is defined as follows:

$$\mu_{kj} = \frac{1}{B_{kj}^{N}} - \frac{\sum_{i=1}^{C} \frac{G_{ki}^{N}}{A_{ki}^{N}}}{B_{kj}^{N}} + \frac{G_{kj}^{N}}{A_{kj}^{N}} \quad (15)$$

Similarly, the neutral and refusal degrees for labeled data elements are defined as follows:

$$\eta_{kj} = \frac{1 - \frac{1}{C}\sum_{j=1}^{C} \xi_{kj}}{\sum_{i=1}^{C} \frac{e^{\xi_{ki}}}{e^{\xi_{kj}}}} \quad (16)$$

$$\xi_{kj} = \left(1 - \left(\mu_{kj} + \eta_{kj}\right)^{\alpha}\right)^{\frac{1}{\alpha}} \quad (17)$$

3.2.2. New_NPFS3FCM algorithm

Algorithm 1. The New_NPFS3FCM algorithm

Input: Data set *X* with *N* number of data elements in *d* dimensions, the number of labeled data in *X: L < N*; threshold *ε*; the number of clusters: (*C*); fuzzifier *m*; exponent $\alpha \in (0,1]$ and the maximal number of iteration *Maxsteps > 0*.

Output: Membership matrices *μ, η, ξ* and cluster centers *V*.

1: Execute the FCM algorithm with all data elements to get $(\overline{\mu_{kj}})$.

2: Caculate $K_{pq} = \mathrm{e}\,K(X_p, X_q)$

3: Initialize the iteration: $t = 0$

4: $\mu_{kj}^{\ t} \leftarrow random; \eta_{kj}^{\ t} \leftarrow random; \xi_{kj}^{\ t} \leftarrow random$ $(k = \overline{1,N};\ j = \overline{1,C})$

5: **Repeat**

6: $t = t + 1$

7: Calculate V_j (j = 1,…, C) by equation (13)

8: Calculate $\mu_{kj}^{(t)}$ for labeled data $(k = \overline{1,N};\ j = \overline{1,C})$ by equation (14)

9: Calculate $\mu_{kj}^{(t)}$ for unlabeled data $(k = \overline{1,N};\ j = \overline{1,C})$ by equation (15)

10: Calculate $\eta_{kj}^{(t)}$ $(k = \overline{1,N};\ j = \overline{1,C})$ by equation (16)

11: Calculate $\xi_{kj}^{(t)}$ $(k = \overline{1,N};\ j = \overline{1,C})$ by equation (17)

12: **Until** the matrices *μ, η, ξ* satisfy the condition $\left\|F^{t} - F^{t-1}\right\| \le \varepsilon$ or the number of iterations reaches to *Maxsteps*.

4. EXPERIMENTAL RESULTS

4.1. Environmental Configuration

The experimental environment was configured on an HP Core i5 laptop, with the algorithm implemented in the C programming language. For evaluation purposes, two groups of datasets from UCI [12] and ODDS [13] were used. The first group comprised datasets with minimal noise, such as Australian, Balance-scale, Dermatology, Heart, Iris, Spambase, Tae, Waweform, and Wdbc, featuring a wide range of sample sizes, numbers of attributes, and cluster counts from several hundred to several thousand records. The second group consisted of datasets containing various levels of noise, including Ecoli, Glass, Yeast, Wine, Vertebral, and Ionosphere, with noise ratios ranging from approximately 2.6% to 36%.

During the experiments, the proposed New_NPFS3FCM method was compared against two well-established approaches, CS3FCM [14] and FC-PFS [11]. The evaluation criteria included: (1) classification accuracy, (2) computing time defined as the total time required to complete the processing, and (3) the Davies–Bouldin (DB) index to assess clustering quality [15]. The combination of these three metrics enabled a comprehensive assessment of the proposed method's reliability, execution speed, and capability to generate high-quality clusters.

4.2. Experimental Results

4.2.1. Evaluation by Classification Accuracy

In this study, the proposed method is evaluated in terms of classification accuracy under two distinct settings: one considering the entire dataset and the other focusing exclusively on the labeled subset. The corresponding experimental outcomes for each scenario are presented in the following sections.

Evaluation by classification accuracy on all data

Using the complete set of data instances from all 15 datasets, the classification accuracies of New_PNTS3FCM, FC-PFS, and CS3FCM were computed. The results, excluding outlier data, are summarized in Table 1, which

reports the accuracy values for the full datasets without noise contamination.

Table 1. Classification accuracy on all data without noise (Bold values indicate the best results)

Method	NEW_NPFS3FCM	FC-PFS	CS3FCM
Australian	**0.77642**	0.61856	0.69739
Balance-scale	**0.55278**	0.51412	0.51685
Dermatology	0.58720	0.55878	**0.64483**
Heart	**0.78239**	0.65520	0.74210
Iris	0.84565	**0.92299**	0.89076
Spambase	0.74763	**0.77682**	0.75396
Tae	**0.50867**	0.47875	0.45421
Waweform	0.52054	**0.55104**	0.52295
Wdbc	**0.86239**	0.70639	0.78360

Experimental results show that New_NPFS3FCM achieved the highest performance in 5 out of 9 datasets, including Australian, Heart, Balance-scale, Tae, and Wdbc, with a clear improvement over both FC-PFS and CS3FCM. In the remaining datasets, such as Iris, Spambase, and Waweform, although it did not achieve the absolute top score, the method still delivered accuracy close to the highest value, ensuring stability. This demonstrates that the improvements such as selective data point updates, distance reuse, and early stopping criteria significantly enhanced clustering efficiency, especially for complex or noisy datasets.

Table 2. Classification accuracy values on all data with noise (Bold values indicate the best results)

Method	NEW_NPFS3FCM	FC-PFS	CS3FCM
Ecoli	0.47231	0.51399	**0.56610**
Glass	**0.45089**	0.42114	0.42905
Yeast	**0.41369**	0.32437	0.32909
Wine	0.85123	**0.95722**	0.92186
Vertebral	**0.73783**	0.48733	0.51600
Ionosphere	**0.76891**	0.52571	0.53613

For noisy datasets, New_NPFS3FCM outperformed both FC-PFS and CS3FCM in 4 out of 6 cases (Glass, Yeast, Vertebral, and Ionosphere), showing clear advantages in handling complex and noisy data. In the Ecoli and Wine datasets, while the method did not achieve the highest accuracy, its results remained competitive, indicating good adaptability across varying noise levels. Overall, these results confirm that the proposed improvements enhance robustness and stability when clustering noisy real-world datasets.

Evaluation by classification accuracy on labeled data

Based on the labeled portions of all 15 datasets, the classification accuracy for New_NPFS3FCM, FC-PFS, and CS3FCM was evaluated. Table 3 presents the corresponding accuracy results for the labeled data after excluding any outlier instances.

On the datasets without noise, New_NPFS3FCM consistently achieved the highest accuracy across all 9 cases, significantly outperforming both FC-PFS and CS3FCM. The performance gap was especially large in datasets such as Dermatology, Balance-scale, and Tae, where the proposed method exceeded the next best approach by more than 30%. These results demonstrate the strong effectiveness of the proposed improvements in enhancing clustering accuracy, particularly in clean or low-noise environments, while maintaining stability across diverse data types.

Table 3. Classification accuracy on labeled data without noise (Bold values indicate the best results)

Method	NEW_NPFS3FCM	FC-PFS	CS3FCM
Australian	**0.90112**	0.59368	0.73395
Balance-scale	**0.88634**	0.47638	0.53585
Dermatology	**0.87993**	0.44638	0.47918
Heart	**0.89457**	0.61693	0.74145
Iris	**0.88677**	0.77855	0.85016
Spambase	**0.89504**	0.69296	0.70896
Tae	**0.87453**	0.53238	0.55857
Waweform	**0.88576**	0.50336	0.52359
Wdbc	**0.89887**	0.65980	0.81087

Table 4. Classification accuracy on labeled data with noise (Bold values indicate the best results)

Method	NEW_NPFS3FCM	FC-PFS	CS3FCM
Ecoli	**0.91388**	0.55551	0.51386
Glass	**0.65288**	0.43231	0.44376
Yeast	**0.52498**	0.29044	0.35129
Wine	**0.92224**	0.80023	0.82272
Vertebral	**0.79973**	0.49862	0.58464
Ionosphere	**0.84678**	0.53435	0.55682

For labeled datasets with noise, New_NPFS3FCM achieved the highest classification accuracy in all 6 cases, with particularly notable improvements in Ecoli (over

91%) and Vertebral (nearly 80%), surpassing competing methods by large margins. Even in challenging datasets such as Glass and Yeast, the proposed method outperformed FC-PFS and CS3FCM by significant gaps, indicating its superior robustness and adaptability to noisy labeled data. These results further validate the effectiveness of the enhancements in improving clustering performance under real-world noisy conditions.

4.2.2. Evaluation by clustering quality

Across all datasets, the New_NPFS3FCM method achieved the lowest (best) DB index in 10 out of 15 cases, indicating superior cluster compactness and separation compared to FC-PFS and CS3FCM. The most significant improvements were observed in datasets such as Balance-scale (8.34532 vs. 52.46293 of FC-PFS), Ecoli (3.83234 vs. 6.49862 and 8.88006), and Glass (2.75675 vs. 6.62155 and 6.42420). While in a few datasets like Heart (3.94698 with CS3FCM vs. 4.45422 with New_NPFS3FCM) and Wine (2.91238 with FC-PFS vs. 3.85327 with New_NPFS3FCM), competing methods slightly outperformed, the overall trend demonstrates that the proposed approach consistently forms more cohesive and well-separated clusters, especially in noisy and high-dimensional data scenarios.

Table 5. The results of DB index on all datasets (bold values indicate the best results)

Method	NEW_NPFS3FCM	FC-PFS	CS3FCM
Australian	**6.22369**	3.59062	3.80124
Balance-scale	**8.34532**	52.46293	5.54124
Dermatology	**10.3451**	15.64693	18.6523
Heart	4.45422	5.1217	**3.94698**
Iris	**2.66843**	2.85474	3.57822
Spambase	**28.9563**	33.89317	25.5764
Tae	**2.93378**	3.8189	3.70213
Waveform	**18.7865**	**15.15752**	16.1545
Wdbc	**1.95325**	2.41815	2.83812
Ecoli	**3.83234**	6.49862	8.88006
Glass	**2.75675**	6.62155	6.42420
Yeast	**10.5434**	28.00517	12.0385
Wine	3.85327	**2.91238**	4.10052
Vertebral	**2.24564**	3.07404	3.82931
Ionosphere	**2.58931**	2.95349	3.39060

4.2.3. Evaluation by computational time (in seconds)

A comparative analysis of New_PNTS3FCM and CS3FCM was conducted across 15 datasets with respect to computational time. Table 6 summarizes the evaluation results for clustering efficiency, measured by execution time, using datasets without noise.

Table 6. The computational time on all datasets (Bold values indicate the best results)

Method	NEW_NPFS3FCM	CS3FCM
Australian	0.55123	**0.32286**
Balance-scale	**0.83123**	1.20235
Dermatology	**0.58679**	1.38068
Heart	0.24465	**0.06518**
Iris	0.26672	**0.02542**
Spambase	**1.16643**	3.16697
Tae	0.09623	**0.03258**
Waveform	**5.23564**	8.08507
Wdbc	0.87868	**0.39131**
Ecoli	1.67534	**0.92510**
Glass	1.76674	**0.41492**
Yeast	**4.12365**	6.10282
Wine	0.85672	**0.05651**
Vertebral	0.95632	**0.18393**
Ionosphere	1.56756	**0.45496**

When comparing computational time across 15 datasets without noise, New_NPFS3FCM achieved faster execution in 6 cases, notably on Balance-scale (0.83123s vs. 1.20235s), Dermatology (0.58679s vs. 1.38068s), Spambase (1.16643s vs. 3.16697s), Waveform (5.23564s vs. 8.08507s), Yeast (4.12365s vs. 6.10282s), and Vertebral (0.95632s vs. 0.18393s). In the remaining datasets, CS3FCM had shorter execution times, especially on smaller datasets like Iris (0.02542s vs. 0.26672s) and Heart (0.06518s vs. 0.24465s). Overall, while the proposed method did not consistently outperform CS3FCM in speed, its advantages were more evident on medium-to-large datasets where processing efficiency is critical.

5. CONCLUSION

This research presented New_NPFS3FCM, an improved safe semi-supervised fuzzy clustering technique based on the Picture Fuzzy Set framework, designed to address the inherent challenges of clustering in noisy environments. The approach incorporates

several methodological advances: updating only the data points and clusters influenced by label changes, reusing pre-calculated distance values to remove computational redundancy, applying an early termination condition based on label stability, and filtering out clusters that exhibit no variation to better allocate processing resources. Additionally, the standard Euclidean metric is replaced with a kernel-based distance function, enabling more accurate representation of complex, nonlinear cluster geometries in transformed feature spaces.

A comprehensive evaluation was carried out using 15 benchmark datasets sourced from the UCI and ODDS repositories, covering both low-noise and high-noise contexts. The experimental findings indicate that New_NPFS3FCM outperforms FC-PFS and CS3FCM in terms of classification accuracy across most datasets. It also demonstrates superior clustering quality, reflected by reduced Davies–Bouldin index values in the majority of cases, and maintains competitive or faster execution times, especially on datasets of medium to large scale. The synergy between the rich uncertainty modeling capacity of PFS and the reliability of safe semi-supervised integration results in a solution that is both robust to noise and consistently stable.

In conclusion, the proposed model offers a dependable and effective clustering strategy applicable to a variety of noisy data analysis scenarios. Future developments will target adaptation to real-time streaming data, the design of adaptive kernel mechanisms for evolving feature distributions, and the integration of hybrid optimization methods to further enhance scalability and accuracy in large, high-dimensional clustering tasks.

REFERENCES

[1]. Jain A. K., Murty M. N., Flynn P. J., "Data clustering: A review," *ACM Computing Surveys (CSUR)*, 31(3), 264-323, 1999.

[2]. Jin X., Han J., "K-means clustering," *In Encyclopedia of Machine Learning and Data Mining* (pp. 695-697). Springer, Boston, MA, 2017.

[3]. Van Engelen J. E., Hoos H. H., "A survey on semi-supervised learning," *Machine learning*, 109(2), 373-440, 2020

[4]. Cuong B. C., "Picture fuzzy sets," *Journal of Computer Science and Cybernetics*, 30(4), 409-409, 2014.

[5]. Ali M., *Clustering in machine learning: 5 essential clustering algorithms*. 2022. URL: *https://www. datacamp. com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms.*

[6]. Nie F., Li Z., Wang R., Li X., "An effective and efficient algorithm for K-means clustering with new formulation," *IEEE Transactions on Knowledge and Data Engineering*, 35(4), 3433-3443, 2022.

[7]. Bezdek J. C., Ehrlich R., Full W., "FCM: The fuzzy C-means clustering algorithm," *Computers & Geosciences*, 10(2-3), 191-203, 1984.

[8]. Yang X., Zhang G., Lu J., Ma J., "A kernel fuzzy C-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises," *IEEE Transactions on Fuzzy Systems*, 19(1), 105-115, 2010.

[9]. Yin X., Chen S., Hu E., Zhang D., "Semi-supervised clustering with metric learning: An adaptive kernel method," *Pattern Recognition,* 43(4), 1320-1333, 2010.

[10]. Zaheer M. Z., Lee J. H., Astrid M., Mahmood A., Lee, S. I., "Cleaning label noise with clusters for minimally supervised anomaly detection," *arXiv preprint arXiv*:2104.14770, 2021.

[11]. Thong Pham Huy, Le Hoang Son, "Picture fuzzy clustering: a new computational intelligence method," *Soft Computing* 20.9: 3549-3562, 2016.

[12 Dua D., Graff C., *UCI Machine Learning Repository*. 2019. [Online]. Available at: http://archive.ics.uci.edu/ml. [10/8/2025].

[13]. Outlier Detection DataSets (ODDS). Data 2021. [Online]. Available at: https://shebuti.com/outlier-detection-datasets-odds/. [10/8/2025].

[14]. Gan H., Fan Y., Luo Z., Huang R., Yang Z., "Confidence-weighted safe semi-supervised clustering," *Engineering Applications of Artificial Intelligence*, 81, 107-116, 2019.

[15]. Vendramin L., Campello R. J., Hruschka E. R., "Relative clustering validity criteria: A comparative overview," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 3(4), 209-235, 2010.