Malaysian Journal of Mathematical Sciences 20(1): 395–421 (2026)
https://doi.org/10.47836/mjms.20.1.19

Malaysian Journal of Mathematical Sciences

Journal homepage: https://mjms.upm.edu.my

A Conceptual of Merging of Intuitionistic Fuzzy C-Means with Chebyshev for Genomic Clustering Solutions Addressing Cancer Issues

Zamri, N. * [1], Bakar, N. A. A. [1], Aziz, A. Z. A. [1], Madi, E. N. [1], Ramli, R. A. [2], Sukono [3], Koon, C. S. [4], and Marhadi [5]

[1]*Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Besut Campus, 22200 Besut, Terengganu, Malaysia*
[2]*Faculty of Medicine, Universiti Sultan Zainal Abidin, Medical Campus, 20400 Kuala Terengganu, Terengganu, Malaysia*
[3]*Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jalan Raya Bandung Sumedang KM21, Jatinagor Sumedang 45363, Jawa Barat Indonesia*
[4]*Jabatan Psikiatri & Kesihatan Mental, Hospital Kuala Lumpur, Jalan Pahang, 50586 Kuala Lumpur, Malaysia*
[5]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Riau, Kota Pekanbaru 28292, Indonesia*

E-mail: nadiahzamri@unisza.edu.my
**Corresponding author*

Received: 18 May 2025
Accepted: 8 September 2025

Abstract

Clustering is a fundamental technique for identifying structures within datasets, with Fuzzy C-Means (FCM) being widely used due to its simplicity and ease of implementation. However, FCM suffers from sensitivity to noise, outliers, and initialization issues. This study introduces an enhanced model, IFCM with Chebyshev, which integrates fuzzy Chebyshev distance and intuitionistic fuzzy sets. Data are first normalized using MinMax scaling, and dimensionality reduction is applied to handle high-dimensional datasets. The optimal cluster number is determined using the Elbow method. The proposed algorithm is evaluated against standard FCM, FCM with Chebyshev, and IFCM. A genomic dataset related to prostate cancer is used as a numerical example. Results show that IFCM with Chebyshev achieves the highest clustering accuracy (88.9%), outperforming IFCM (85.7%), FCM with Chebyshev (81.2%), and FCM (78.5%). It also yields superior cluster validity indices, recording the highest Partition Coefficient (0.74) and the lowest Partition Entropy (0.52), indicating clearer cluster separation. Although IFCM with Chebyshev incurs higher computational cost (1.58s), sensitivity analysis demonstrates faster convergence around five clusters, suggesting an optimal structure. Memory consumption remains consistent at 295 MB across cluster settings, highlighting efficiency for large-scale applications. Overall, combining Chebyshev distance and IFCM enhances clustering robustness and accuracy, particularly in noisy or complex data environments, making it a promising approach for advanced data analysis tasks.

Keywords: clustering; FCM; intuitionistic FCM; Chebyshev; prostate cancer; genomic clustering.

1 Introduction

In today's fast-paced, technology-driven world, innovation is continually advancing, with machine learning emerging as a prominent development within this digital revolution. Machine learning involves the creation of computer models capable of learning and making autonomous predictions or decisions based on provided data [40]. As technology progresses, machine learning has become a transformative force, redefining how we process information, solve complex problems, and make critical decisions. By harnessing data-driven algorithms, machine learning has unlocked unparalleled potential across various industries and applications. As a subset of artificial intelligence, machine learning has revolutionized how computers analyze data and make decisions [20]. It enables systems to learn from data, detect hidden patterns, and improve their performance over time without requiring explicit programming. This adaptability has led to widespread adoption in fields such as image recognition [36], natural language processing [3], and medical diagnostics [16]. The core methodologies of machine learning are categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning [10]. Supervised learning [5] involves training algorithms on labeled datasets to predict outcomes based on input-output relationships. In contrast, unsupervised learning [9] focuses on identifying patterns and structures in unlabeled data without predefined guidance. Reinforcement learning [26], on the other hand, relies on agents interacting with their environment, learning from rewards and penalties to maximize cumulative rewards. Each machine learning category offers unique approaches and serves distinct purposes, contributing to advancements in prediction accuracy [37], pattern discovery [23], and decision-making optimization [6]. These diverse paradigms underscore the dynamic and ever-evolving nature of machine learning in addressing real-world challenges.

Machine learning can be broadly classified into three main categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning entails training algorithms on labeled datasets, enabling them to predict outcomes based on established input-output relationships. In contrast, unsupervised learning deals with unlabeled data, emphasizing the discovery of patterns and structures within the data. Reinforcement learning, on the other hand, involves an agent that learns by interacting with its environment, using feedback to maximize rewards over time. Each of these approaches plays a crucial role in diverse applications, enhancing predictive accuracy, uncovering hidden patterns, and optimizing decision-making processes, showcasing the versatility and depth of machine learning.

Our research emphasizes unsupervised learning, a rapidly advancing field celebrated for its ability to uncover meaningful insights from unlabeled data. Clustering methods, a cornerstone of unsupervised learning, are instrumental in grouping similar data objects based on their inherent features. Among these methods, the K-means algorithm is widely recognized [18]. It partitions data into k unique clusters, each represented by a centroid, and is extensively applied in image segmentation [28], customer segmentation [32], and document clustering [22]. C-Means Clustering, often synonymous with K-Means Clustering, is a popular unsupervised machine learning algorithm used to segment datasets into distinct clusters based on feature similarities. While it is lauded for its simplicity and efficiency, the conventional C-Means algorithm encounters challenges in handling the uncertainty and imprecision present in real-world datasets, especially in complex domains like genomics. FCM [12] is a noteworthy extension of K-means, allowing data points to belong to multiple clusters with varying degrees of membership. This flexibility makes FCM particularly effective in scenarios involving ambiguous memberships, with applications in pattern recognition, medical imaging, and fuzzy control systems. Further enhancing this approach, Intuitionistic Fuzzy C-Means (IFCM) builds upon FCM by introducing a hesitation degree, which captures the uncertainty in data point memberships. This capability enables IFCM to manage ambiguous data points more effectively, making it suitable for handling complex clustering challenges.

Despite their effectiveness, clustering algorithms like K-means, FCM, and IFCM rely heavily on distance measures, such as the Euclidean distance, to assess the similarity between data points. While Euclidean distance is a common metric, it faces significant limitations in high-dimensional spaces, where it becomes susceptible to issues associated with the "curse of dimensionality" [25]. This challenge underscores the need for innovative approaches to improve clustering performance in complex datasets.

Therefore, this study aims to integrate the IFCM algorithm with the Chebyshev distance measure as an alternative to the traditional Euclidean distance in clustering tasks. The Chebyshev distance [8], also referred to as chessboard distance or L^{∞} distance, determines the maximum absolute difference between corresponding elements of two data points [15]. It is particularly well-suited for scenarios where movement is restricted to a grid-like structure and demonstrates greater resilience to outliers compared to Euclidean distance. One of the key advantages of Chebyshev distance lies in its simplicity and computational efficiency, making it an attractive choice for high-dimensional datasets and real-time applications. Its straightforward calculation requires minimal computational resources, ensuring scalability for large datasets. Additionally, the robustness of the Chebyshev distance minimizes the influence of outliers, preventing extreme values in individual dimensions from disproportionately affecting clustering outcomes. This property enhances the accuracy and reliability of data groupings, making the Chebyshev distance a valuable tool for improving clustering performance.

Numerous studies have explored the application of Chebyshev distance in conjunction with FCM. For instance, Zamri et al. [41] introduced a modified FCM algorithm incorporating Fuzzy Chebyshev distance. Their methodology included pre-processing raw datasets with the MinMax Scaler and implementing dimensionality reduction techniques to mitigate issues associated with multidimensional datasets. Suitable numbers of clusters are determined using the Elbow method, while Fuzzy Chebyshev distance replaced the conventional distance metric in the FCM algorithm. A comparative evaluation against existing methods, supported by a numerical example using genomic data from prostate cancer, confirmed the feasibility and comparable performance of the proposed approach. Then, Kumar et al. [17] proposed a hybrid fuzzy clustering technique based on the fusion of an intuitionistic modified fuzzy C-means algorithm and an improved genetic algorithm. Their study addresses the problem of high sensitivity to initial centroids by employing a genetic algorithm enhanced with a normalized crossover and mutation operator. Moreover, their proposed algorithm reduces the effect of noise by developing a new metric and resolves uncertainty in assigning membership values through the use of two negation functions.

Kumar and Kumar [16] improved the FCM and particle swarm optimization (PSO) clustering technique to handle the initialization problem. According to their study, PSO effectively enhances the performance of the improved FCM, thereby increasing clustering effectiveness. Their results show that the proposed algorithm produces encouraging outcomes compared to established clustering techniques in the literature. Similarly, Rusdiana et al. [19] investigated optimal distance metrics for FCM clustering. Their study evaluated various distance measures, including Euclidean, Manhattan, Chebyshev, and Minkowski distances, using metrics such as the partition coefficient index (PC), modified partition coefficient index (MPC), and RMSE. The findings indicated that Manhattan distance was optimal for datasets with two clusters, while Minkowski distance was superior for datasets with three clusters, highlighting the effectiveness of these measures for FCM in different scenarios. Supianto et al. [29] improved an FCM clustering performance using PSO on student grouping based on learning activity in a digital learning media. Result showed that the adaption of FCM and PSO improves the clustering quality significantly based on the observed average silhouette coefficient.

Javadian et al. [14] proposed a re-fuzzification algorithm derived from FCM, based on the

shape and density of clusters. They tested their proposed method using simulations on both real and synthetic datasets. Their findings indicate that the re-fuzzification algorithm can slightly improve the clustering quality of FCM, as data points are reassigned based on similarity to the shape and density of their respective clusters. Additionally, Surono and Putri [30] proposed an objective function for FCM that combined Minkowski and Chebyshev distances, utilizing principal component analysis (PCA) for dimensionality reduction. Their results revealed improved clustering accuracy, with a minimum objective function value of 0.0373 achieved in the 15^{th} iteration, out of a maximum of 100 iterations. To date, however, no studies have specifically examined the integration of IFCM with Chebyshev distance, leaving a gap in the current research landscape.

To illustrate the practical application of Chebyshev distance and unsupervised learning in a real-world context, this study presents a case study on cancer cell clustering. In medical research and diagnostics, classifying genomic cancer cells based on features such as single nucleotide variants, insertions or deletions, and genomic rearrangements is of critical importance [11]. This case study applies FCM and IFCM algorithms, incorporating Chebyshev distance, to analyze a dataset of prostate cancer cell genomic features. By identifying clusters of genomic cancer cells with similar characteristics, this approach aims to provide valuable insights into the diverse subtypes of cancer, offering potential implications for prognosis and treatment strategies. Integrating Chebyshev distance into these algorithms seeks to improve their accuracy and robustness, addressing limitations associated with traditional measures like Euclidean distance. Additionally, its simplicity, resilience to outliers, and suitability for binary data are expected to reduce computational time, making it an efficient choice for various clustering tasks. This research aspires to uncover meaningful insights that can significantly advance medical research, particularly in cancer diagnosis and treatment, while demonstrating the potential of enhanced clustering methodologies in addressing complex medical challenges.

2 Theoretical Background

This section provides an overview of the fundamental concepts related to FCM and Fuzzy Chebyshev techniques.

2.1 Fuzzy C-Means (FCM)

FCM is a commonly used fuzzy clustering algorithm known for its simplicity and ease of implementation. It assigns data points to multiple clusters based on membership values [16]. Extensive research has been conducted on the FCM algorithm. For example, Eryoldaş and Durmuşoğlu [10] proposed a Latin Hypercube Hammersley Sampling (LHHS) based integrating features from the Artificial Bee Colony (ABC) and Standard Genetic Algorithm (SGA) to optimize training budget configurations. Chen et al. [5] applied phase-space reconstruction with FCM Clustering for predicting and classifying ventricular arrhythmia. Meanwhile, Elshenawy et al. [9] introduced a fault detection and diagnosis approach that combines k-nearest neighbors with FCM Clustering. They integrated k-nearest neighbors and FCM Clustering for error recognition and evaluation. Shi et al. [26] conducted a contrastive evaluation using multiscale residual U-Net and FCM Clustering for pulmonary nodules segmentation. Although many studies have explored modifications to the algorithm, it continues to face challenges with high uncertainty, impacting clustering performance [37]. Additionally, it often converges to locally optimal solutions. To address these challenges, further modifications are necessary.

FCM Clustering minimizes the objective function by iteratively updating the clustering centers and fuzzy membership degrees. In this process, samples are assigned to various categories based on the principle of maximum membership. The optimization model for the FCM algorithm is described as follows:

$$\min F_{\text{FCM}}(U,V)=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}d^2(x_i,v_j), \quad \text{s.t.} \quad \begin{aligned} &0\le u_{ij}\le 1, && 1\le i\le n,\ 1\le j\le c,\\ &\sum_{j=1}^{c}u_{ij}=1, && 1\le i\le n,\\ &0<\sum_{i=1}^{n}u_{ij}<n, && 1\le j\le c, \end{aligned} \tag{1}$$

where u_{ij} represents the membership degree of data point x_i in cluster j; v_j is the centroid of cluster j; $d^2(x_i,v_j)$ denotes the squared Euclidean distance between x_i and v_j as $||x_i-v_j||^2$; and $m\ge 1$ is fuzzy weighting exponent, and is usually set as 2; c is the number of clusters; n represents the number of samples.

By Langrange multiplier method, the expressions of fuzzy membership degree u_{ij} and the clustering center $v_j$are obtained as follows:

$$u_{ij}=\frac{1}{\sum_{r=1}^{c}\left(\frac{d^2(x_i,v_j)}{d^2(x_i,v_r)}\right)^{\frac{1}{m-1}}}, \tag{2}$$

while cluster centroids are recalculated as:

$$v_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}x_i}{\sum_{i=1}^{n}u_{ij}^{m}}. \tag{3}$$

The FCM algorithm is a powerful clustering method widely applied in IoT, machine learning, bioinformatics and etc.

2.2 Intuitionistic Fuzzy C-Means (IFCM)

To address the limitations of fuzzy sets, Atanassov [1] expanded upon Zadeh's [40] concept of fuzzy sets by developing the intuitionistic fuzzy set (IFS). This new framework introduces two additional components: non-membership (non-belonging) and hesitation. According to Atanassov [2], the IFS can be defined as follows:

Let $X=\{x_1,x_2,\ldots,x_n\}$ be a collection of elements x. Then, IFS be set A for X can be denoted as:

$$A=\{\langle x,\mu_A(x),\nu_A(x)\rangle:\forall\ x\in X\}, \tag{4}$$

where $\mu_A:X\to[0,1]$ is membership degree and $\nu_A:X\to[0,1]$ is non-membership degree for all elements of $x\in X$ to the set A with the condition that membership degree and non-membership degree satisfy $0\le\mu_A(x)+\nu_A(x)\le 1$. Therefore, the hesitation degree can be calculated by

$$\pi_A(x)=1-\mu_A(x)-\nu_A(x), \tag{5}$$

with $\pi_A : X \to [0, 1]$.

From the definition of IFS, Xu and Wu [39] utilized the IFS into the FCM algorithm. The new modification of objective function for IFCM can be shown as:

$$J(U, X, G) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij}^{f} \|X_i - G_j\|^2, \tag{6}$$

In (6), k is the number of clusters, n is the number of data, u_{ij} is the degree function of matrix U in IFS as can be calculated as:

$$u_{ij} = \frac{1}{\sum_{r=1}^{k} \left(\frac{\|X_i - G_j\|^2}{\|X_i - G_r\|^2} \right)^{2/(f-1)}}, \tag{7}$$

Meanwhile, f is fuzziness parameter, X_i is the data in term of the intuitionistic fuzzy set as $X_i = (\mu(x_i), \nu(x_i), \pi(x_i))$, G_j is the clusters centroid in term of intuitionistic fuzzy set as $G_j = (\mu(g_j), \nu(g_j), \pi(g_j))$ is defined as:

$$\mu(g_j) = \frac{\sum_{i=1}^{n} u_{ij}^{f} \mu(x_i)}{\sum_{i=1}^{n} u_{ij}^{f}}, \tag{8}$$

$$\nu(g_j) = \frac{\sum_{i=1}^{n} u_{ij}^{f} \nu(x_i)}{\sum_{i=1}^{n} u_{ij}^{f}}, \tag{9}$$

$$\pi(g_j) = \frac{\sum_{i=1}^{n} u_{ij}^{f} \pi(x_i)}{\sum_{i=1}^{n} u_{ij}^{f}}, \tag{10}$$

The term $\|X_i - G_j\|$ in (6) and (7) are the Euclidean distance measure between the data X_i and the centroids G_j that are elaborated as in [31],

$$d^2(X_i, G_j) = \|X_i - G_j\| = (\mu(x_i) - \mu(g_j))^2 + (\nu(x_i) - \nu(g_j))^2 + (\pi(x_i) - \pi(g_j))^2. \tag{11}$$

With iterative updates of degree function matrix U and centroids G_j, the IFCM iteratively optimizes the objective function $J(U, X, G)$ in (6) until $|U^{(\text{new})} - U^{(\text{old})}| \leq \epsilon$, where ϵ is the termination criterion.

2.3 Chebyshev distance

In this study, our goal is to enhance unsupervised machine learning for clustering by modifying the traditional distance measure used in k-means, FCM, and IFCM algorithms. Rather than relying

on the Euclidean distance measure, we propose a new clustering approach that incorporates the Chebyshev distance measure. The formula for Chebyshev distance is defined by Kim and Mueller [15] as follows:

$$d(x_i, C_j) = \max |x_i - C_j| . \tag{12}$$

Based on the findings and limitations highlighted in the literature review, this study introduces an integrated approach that bridges these gaps by incorporating IFCM with Chebyshev. Section 3 presents an in-depth review of the proposed approach, outlining the sequential process along with the relevant mathematical formulations.

3 The Integration of IFCM with Chebyshev Distance

This section describes the stages of the framework, as depicted in Figure 1, which illustrates the fundamental process of the study. The framework is logically structured into four essential stages. Phase 1 involves dataset preparation and pre-processing. Phase 2 examines techniques for dimensionality reduction. Phase 3 is dedicated to identifying the best number of groupings. Finally, Phases 4a and 4b introduces the proposed approach through a systematically organized step-by-step process. Each phase is thoroughly explained to provide a clear understanding of the processes involved.

Phase 1: *Dataset Preparation and Pre-processing*

Our dataset comprises multiple modalities due to the diverse feature types it includes, such as binary attributes, real-valued data, and Sequence Ontology (SO) terms. Therefore, data pre-processing is a crucial step before implementing machine learning algorithms. For SO terms, a quantitative rating system can be established considering their classification levels: Minimal, Slight, Medium, and Significant. These scores, which estimate the impact of each variant, can be derived using the Ensembl Variation – Calculated Variant Consequences and subsequently applied to gene mutations. Following this, data normalization is conducted across the dataset using both Standard Scaler and MinMax Scaler. The Standard Scaler standardizes the data by subtracting the mean and scaling it to unit variance, while the MinMax Scaler operates by normalizing each feature by subtracting the minimum value and dividing by the range, ensuring consistency across different data types.

Phase 2: *Dimensionality Reduction*

Our datasets typically contain a high number of features relative to the available samples, necessitating dimensionality reduction before clustering. High-dimensional data can negatively impact the effectiveness of many statistical methods, a challenge known as the "curse of dimensionality" [18]. Various techniques are available for dimensionality reduction, each differing in interpretability and computational complexity. Principal Component Analysis (PCA) is frequently employed for this purpose due to its computational efficiency and ability to improve machine learning performance on high-dimensional data.

***Phase 3**: Determining the Optimal Number of Clusters*

Identifying the ideal number of clusters can be challenging, especially for datasets containing diverse parameter types. Diverse parameter types refer to the variety in data types, scales, distributions, and quality across the features in a dataset. This diversity makes clustering more challenging and highlights the importance of proper preprocessing, parameter tuning, and distance metric selection to ensure that meaningful and accurate clusters are formed. Next, estimating the most suitable number of clusters is a crucial aspect of clustering algorithms. To address this, the Elbow method is employed. Alongside Silhouette analysis, this method involves plotting the explained variance against the number of clusters and selecting the inflection point where the curve forms an "elbow," indicating the optimal number of clusters.

***Phase 4a**: Proposed FCM with Fuzzy Chebyshev*

The FCM algorithm can be modified using the following approach:

Step 1: Define the best value for k, f, and ϵ
Let $X = \{x_1, x_2, \ldots, x_n\}$ where $x_i \in X$ is the data elements, the modified objective function for FCM is achieved as:

$$J^c(x_i, C_j) = \sum_{j=1}^{k}\sum_{i=1}^{n}(\mu_{ij})^f \left(\max|x_i - C_j|\right)^2, \quad 1 \leq f \leq \infty, \tag{13}$$

where k represents the total number of clusters, n denotes the number of data points, C_j is the centroid of the cluster, $\mu_{ij} \in [0, 1]$ signifies the membership degree of the data point x_i in cluster C_j, and f is the fuzziness parameter that defines the degree of association between x_i and the clusters.

Step 2: Assign an initial random membership value
Equations (14)-(15) adheres to the following conditions:

$$\sum_{j=1}^{k}\mu_{ij} = 1, \ \forall\ j, \tag{14}$$

$$0 < \sum_{i=1}^{n}\mu_{ij} < n, \ \forall\ i. \tag{15}$$

Step 3: Compute centroids (C_j)

$$C_j = \frac{\sum_{i=1}^{n}(\mu_{ij})^f x_i}{\sum_{i=1}^{n}(\mu_{ij})^f}. \tag{16}$$

Step 4: Recalculate the membership value μ_{ij} utilizing the Chebyshev distance measure. Additionally, the μ_{ij} formula requires modification as follows:

$$\mu_{ij} = \frac{1}{\sum_{r=1}^{k}\left(\frac{\max|x_i - C_j|}{\max|x_i - C_r|}\right)^{2/(f-1)}}.$$

Step 5: Continue iterating between Steps 3 and 4 until the condition $\left|J^{C^{\text{new}}} - J^{C^{\text{old}}}\right| \leq \epsilon$ is met or the maximum number of iterations is reached. During the iteration process, the objective function is evaluated until the error value ϵ falls within the desired threshold.

The modified FCM algorithm can be summarized as follows:

Algorithm 1: Proposed FCM with Fuzzy Chebyshev
Step 1: Define the best value for k, f, and ϵ.
Step 2: Assign an initial random membership value.
Step 3: Compute the centroids C_j.
Step 4: Recalculate the membership value μ_{ij} utilizing the Chebyshev distance measure.
Step 5: Repeat Steps 3 and 4 until the condition $\left|J^{C^{\text{new}}} - J^{C^{\text{old}}}\right| \leq \epsilon$ is met or the maximum number of iterations is reached.

***Phase 4b**: Proposed IFCM with Fuzzy Chebyshev*

For IFCM, the clustering data must first be converted into intuitionistic fuzzification data and an intuitionistic fuzzy generator before the IFCM may be modified. Thus, the fuzzification data can be denoted as:

$$\mu(x_i) = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}}, \tag{17}$$

where $(x_i)_{\max}$ and $(x_i)_{\min}$ are the maximum and minimum values of x_i, respectively. Then, in order to generate the non-membership degree of the data, we use the following formula:

$$\nu(x_i) = \frac{1 - \mu(x_i)}{1 + \lambda\mu(x_i)}, \quad \lambda > 0. \tag{18}$$

Therefore, the collection of data in an Intuitionistic Fuzzy Set (IFS) $X = \{x_1, x_2, \ldots, x_n\}$ can be shown as:

$$B = \{\langle x, \mu_B(x), \nu_B(x)\rangle \mid \forall\, x \in X\}, \tag{19}$$

where $\mu_B : X \to [0,1]$ is the membership degree and $\nu_B : X \to [0,1]$ is the non-membership degree for all elements $x \in X$ to the set B satisfy with
$0 \leq \mu_B(x) + \nu_B(x) \leq 1$, the condition that membership degree and non-membership degree. The hesitation degree can be computed by

$$\pi_B(x) = 1 - \mu_B(x) - \nu_B(x), \tag{20}$$

where $\pi_B : X \to [0,1]$.

By applying the Chebyshev distance measure, the new objective function of the Improved Fuzzy C-Means (IFCM) can be defined as:

$$J^C(\tilde{U}, X, G) = \sum_{j=1}^{k}\sum_{i=1}^{n} \tilde{u}_{ij}^{f} \left(\max|X_i - G_j|\right)^2, \tag{21}$$

where f is the fuzziness parameter, k is the number of clusters, and n is the number of data points. The modified degree function of matrix $\tilde{U}$ is

$$\tilde{u}_{ij} = \frac{1}{\sum_{r=1}^{k} \left(\frac{\max |X_i - G_j|^2}{\max |X_i - G_r|^2} \right)^{\frac{2}{f-1}}}, \tag{22}$$

where X_i is represented as $X_i = (\mu(x_i), \nu(x_i), \pi(x_i))$. G_j is the clusters centroid as $G_j = (\mu(g_j), \nu(g_j), \pi(g_j))$ which is obtained using,

$$\mu(g_j) = \frac{\sum_{i=1}^{n} \tilde{u}_{ij}^{f} \mu(x_i)}{\sum_{i=1}^{n} \tilde{u}_{ij}^{f}}, \tag{23}$$

$$\nu(g_j) = \frac{\sum_{i=1}^{n} \tilde{u}_{ij}^{f} \nu(x_i)}{\sum_{i=1}^{n} \tilde{u}_{ij}^{f}}, \tag{24}$$

$$\pi(g_j) = \frac{\sum_{i=1}^{n} \tilde{u}_{ij}^{f} \pi(x_i)}{\sum_{i=1}^{n} \tilde{u}_{ij}^{f}}. \tag{25}$$

The modified distance $\max |X_i - G_j|$ in (21) and (22) is the Chebyshev distance measure between the data X_i and the centroids G_j, which can be elaborated as:

$$\max |X_i - G_j| = \max(|\mu(x_i) - \mu(g_j)|, |\nu(x_i) - \nu(g_j)|, |\pi(x_i) - \pi(g_j)|). \tag{26}$$

With iterative updates of the degree function matrix $\tilde{U}$ and centroids G_j, the modified IFCM iteratively optimizes the objective function $J^C(\tilde{U}, X, G)$ until $|\tilde{U}^{\text{new}} - \tilde{U}^{\text{old}}| \leq \epsilon$ which satisfies the termination criterion ϵ.

Algorithm 2: Enhanced IFCM with Fuzzy Chebyshev
Step 1: Define the number of clusters (k), fuzziness parameter (f), lambda (λ), and stopping criteria (ϵ).
Step 2: Compute $X_i = (\mu(x_i), \nu(x_i), \pi(x_i))$ for $i = 1, 2, \ldots, n$ using (18) and (20).
Step 3: Initialize the fuzzy membership matrix $\tilde{U}$.
Step 4: Determine the centroid values $G_j = (\mu(g_j), \nu(g_j), \pi(g_j))$ using (23)-(25).
Step 5: Update the new fuzzy degree $\tilde{u}_{ij}$ based on the Chebyshev distance metric using (22).
Step 6: Iterate Steps 4 and 5 until $|\tilde{U}^{\text{new}} - \tilde{U}^{\text{old}}| \leq \epsilon$ or $|J^{C^{\text{new}}} - J^{C^{\text{old}}}| \leq \epsilon$, ensuring that the objective function $J^C(\tilde{U}, X, G)$ reaches its optimal value.

Hence, Phases 1 to 4b are recapped as depicted in Figure 1.

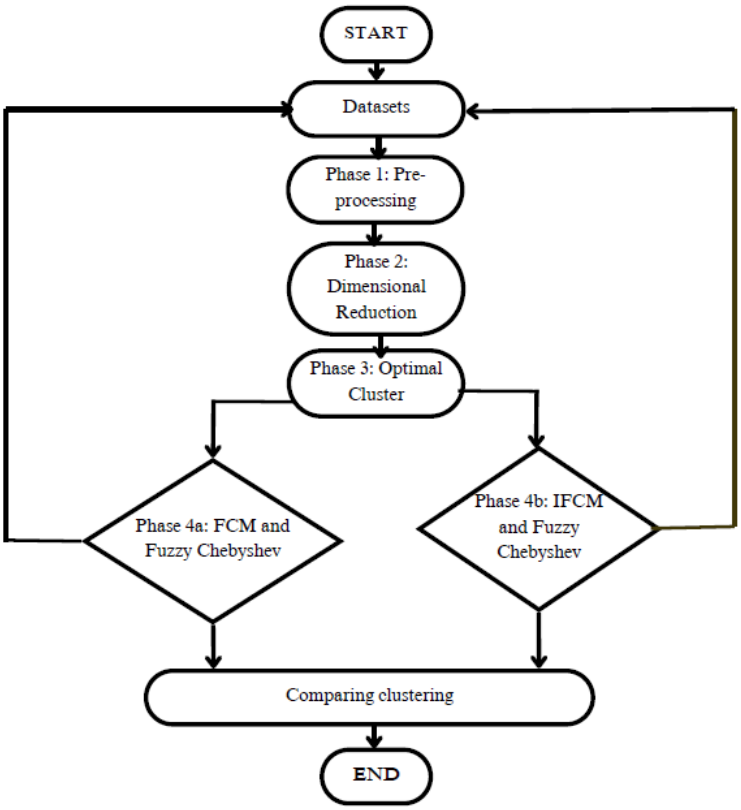


Figure 1: Stages of the proposed approach.

4 Utilization of Genomic Clustering Approaches

Prostate cancer (PCa) is a prevalent malignancy, with around 1.4 million new diagnoses recorded in 2020 [28]. Although the fatality rate of prostate cancer is relatively lower than many other cancers, its high prevalence led to over 375,000 deaths in the same year [28]. This malignancy is marked by its unpredictable course; some patients may undergo swift metastasis resulting in death within a few years, while others may live for decades with confined disease [34]. Given the frequent diagnosis of prostate cancer and the variability in its progression, it is vital to identify disease pathways early to avoid unnecessary treatments, particularly through the analysis of genomic subtypes. Numerous studies have explored different facets of prostate cancer. For example, Salman et al. [22] developed a detection and diagnosis system utilizing an artificial intelligence approach based on the YOLO object detection algorithm. In another study, Hassan et al. [12] classified MRI images of prostate cancer using artificial intelligence methods, including supervised machine learning and deep learning. Likewise, Shaikh and Rao [25] forecast cancer evolution using various machine learning methods such as Decision Trees (DTs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs). Bustamam et al. [4] classified prostate cancer using SVM integrated with Recursive Feature Elimination and a One-Dimensional NaÃ¯ve Bayes Classifier. Despite significant research advancements, the genomic features of prostate cancer remain insufficiently studied. Additionally, limited exploration has been conducted on familial patterns and genetic associations related to prostate cancer, particularly in the context of integrating machine learning with fuzzy mathematics.

This study utilizes prostate cancer data consisting of 839 samples, including all Copy Number Alterations (CNAs), genomic signatures, and driver genes, with 99 unique features adapted from Eagles [8]. The dataset represents diverse ethnic groups, including individuals of European, African, and Asian ancestry. Before analysis, the data underwent pre-processing to enhance the performance of machine learning algorithms, as working with raw data could reduce their effectiveness. The MinMax Scaler was applied to normalize all data types by rescaling feature values between 0 and 1 based on their minimum and maximum values. Additionally, Principal Component Analysis (PCA) was performed on the normalized dataset, which includes 99 attributes, to identify key factors associated with prostate cancer. According to criteria established by Kim and Mueller [15], factors with eigenvalues of 1.0 or greater are considered significant. According to Jackson [13], selecting factors with the highest eigenvalues improves the interpretability of the data structure, as the identified Principal Components (PCs) typically capture more meaningful information than the original variables.

After running PCA, 75.19% of the variance was captured along with 50 principal components (PCs) from the original dataset. Among them, 75 features displayed strong positive loadings, while 24 features had weaker positive loadings. The analysis focused on these 75 significant features, as outlined in Table 1.

Table 1: Cumulative variance of PCs.

PCs	PC1	PC2	PC3	...	PC48	PC49	PC50
Eigenvalue	7.282	3.900	3.482	...	0.87705	0.873	0.859
Variability (%)	7.347	3.935	3.513	...	0.884853	0.881	0.866
Cumulative (%)	7.347	11.282	14.795	...	0.884853	74.325	75.191

Determining the suitable number of clusters is a challenging process, particularly when dealing with various parameters and different types of datasets. Additionally, a crucial aspect of clustering methods is identifying the most suitable number of clusters for the available dataset. To address this problem, we employed three distinct methods: 1) Davies-Bouldin Index, 2) Within-Cluster Sum of Squares (WCSS), and 3) Xie-Beni Index. The Davies-Bouldin Index, first proposed by Davies and Bouldin [7], quantifies the average "similarity" between clusters based on the ratio of inter-cluster distance to intra-cluster size. A model with a lower Davies-Bouldin Index reflects superior differentiation between clusters. The subplot of the Davies-Bouldin Index in Figure 2 indicates that the ideal cluster count is three for the FCM algorithm, when using Chebyshev distance.

The WCSS quantifies the total squared distance from each data point to its cluster centroid. When graphed against varying numbers of clusters (K), the WCSS typically produces a curve that resembles an "elbow." As the number of clusters increases, WCSS values generally decline. Analyzing the WCSS subplot depicted in Figure 3 indicates that the optimal number of clusters for the FCM algorithm, when utilizing the Chebyshev distance, is estimated to be three.

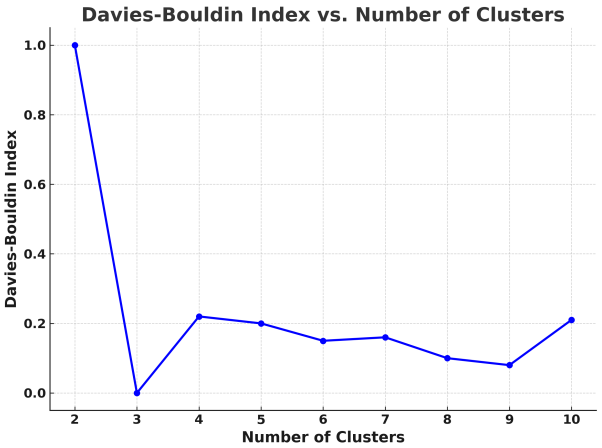


Figure 2: Ideal k based on the Davies-Bouldin Index (FCM with Chebyshev distance).

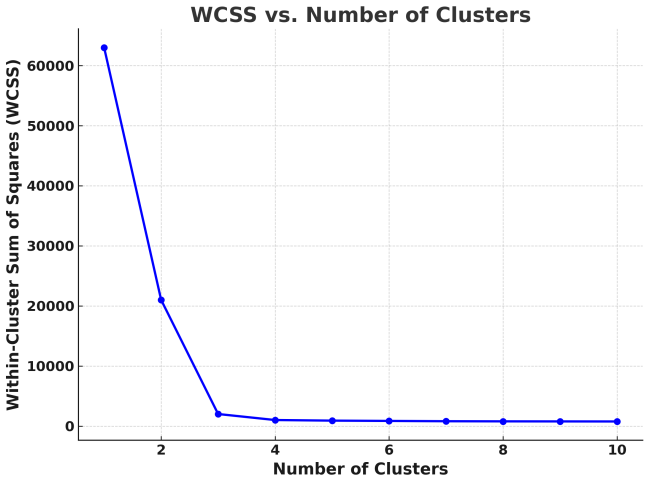


Figure 3: Ideal k based on the WCSS (FCM with Chebyshev).

The Xie-Beni index evaluates how closely packed and distinct clusters are by computing the ratio of the total dispersion within clusters to the smallest dispersion between clusters. The subplot of the Xie-Beni Index in Figure 4 shows that the optimal number of clusters for the FCM algorithm using Chebyshev distance is three.

To find the most suitable number of clusters for the IFCM algorithm using Chebyshev distance, we applied several evaluation metrics, including the Davies-Bouldin index, WCSS, and the Xie-Beni index. These methods involve creating a plot of explained variance as a function of the number of clusters and then identifying the point on the curve that indicates the ideal number of clusters. The analysis of the Davies-Bouldin index subplot for the IFCM with Chebyshev distance suggests that the optimal number of clusters to use is three. The provided figure depicts the Davies-Bouldin Index values for clustering solutions across different numbers of clusters, ranging from 2 to 10. The Davies-Bouldin Index is a metric used to evaluate the quality of clustering, where lower values indicate better-defined clusters with higher separation and cohesion.

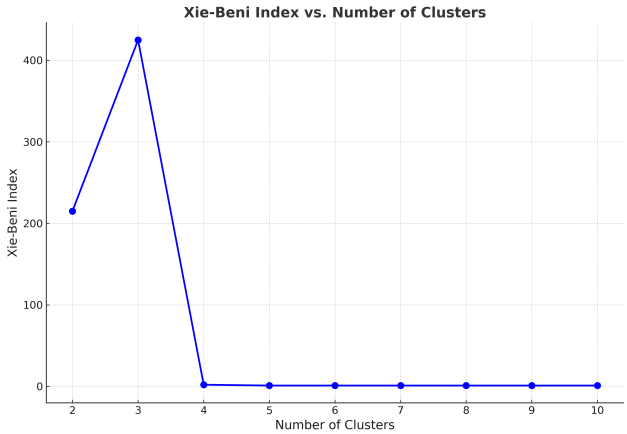


Figure 4: Ideal k based on the Xie-Beni Index (FCM with Chebyshev).

From Figure 5, the Davies-Bouldin Index shows a significant drop when moving from 2 to 3 clusters, indicating an improved clustering solution. This suggests that 3 clusters might provide better-defined and more distinct groupings compared to two clusters. For cluster numbers greater than three, the Davies-Bouldin Index values fluctuate but remain relatively low. This indicates that these clustering solutions are comparable in quality, with no dramatic improvements or deterioration. Based on the DBI values, the clustering solution with three clusters appears to be the most optimal, as it achieves the lowest Davies-Bouldin Index value, signifying well-separated and compact clusters. This analysis demonstrates the utility of the Davies-Bouldin Index in determining the optimal number of clusters for unsupervised learning tasks.

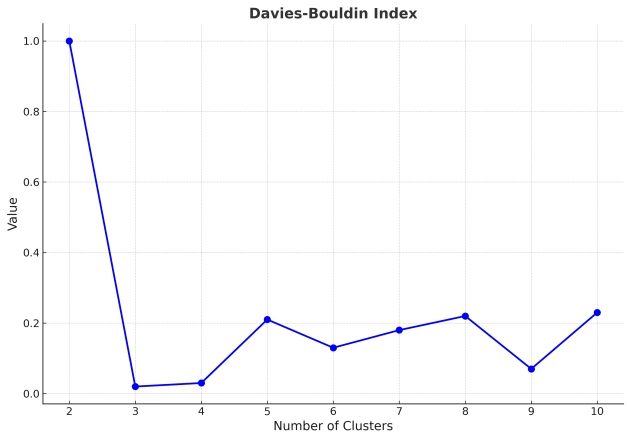


Figure 5: Optimal k using Davies-Bouldin Index (IFCM with Chebyshev).

Next, the provided Figure 6 illustrates the WCSS values for different numbers of clusters, ranging from 2 to 10. WCSS measures the total variance within each cluster and is commonly used to evaluate the compactness of clusters. Lower WCSS values indicate more compact clusters with minimal variance. From Figure 6, there is a steep drop in WCSS as the number of clusters increases from 2 to 4, indicating that increasing the number of clusters significantly reduces the intra-cluster variance and leads to better-defined clusters. Beyond four clusters, the rate of decrease in WCSS slows down considerably, forming an "elbow" in the curve. This suggests that adding more clusters beyond this point results in diminishing returns in reducing WCSS. The el-

bow point is typically used to determine the optimal number of clusters. Based on this figure, four clusters appear to be a reasonable choice for balancing cluster compactness and complexity.

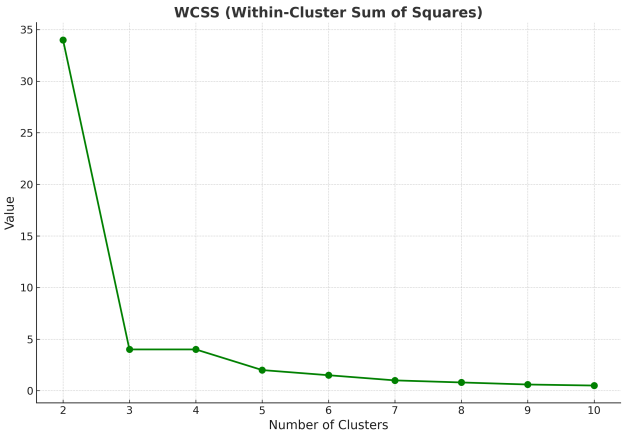


Figure 6: Optimal k using WCSS (IFCM with Chebyshev).

The provided Figure 7 illustrates the Xie-Beni Index values for different numbers of clusters, ranging from 2 to 10. The Xie-Beni Index is used to evaluate the clustering quality by considering both the compactness of clusters (within-cluster variance) and the separation between clusters. Lower values of the Xie-Beni Index indicate better clustering results with more compact and well-separated clusters. There is a sharp decline in the Xie-Beni Index when the number of clusters increases from 2 to 3, suggesting that the clustering solution improves significantly as the number of clusters increases from 2 to 3. Beyond 3 clusters, the Xie-Beni Index remains nearly constant, indicating that increasing the number of clusters further does not result in significant improvement in the clustering quality. The lowest value occurs at 3 clusters, which indicates that the optimal clustering solution, according to the Xie-Beni Index, is 3 clusters. This suggests that 3 clusters offer the best balance of compactness and separation. Overall, based on the Xie-Beni Index, the optimal number of clusters is 3, as this minimizes the index and provides the best clustering solution.

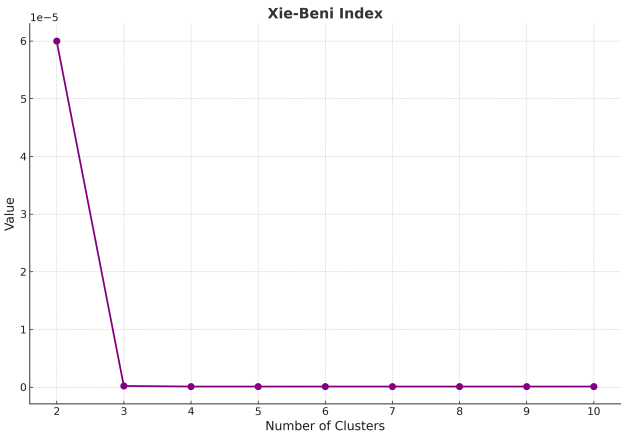


Figure 7: Optimal k using Xie-Beni Index (IFCM with Chebyshev).

Based on the results shown in Table 2, the majority of the data points are assigned to Clus-

ter 1 (73 points). This indicates that, regardless of the algorithm used, the data is predominantly grouped into this cluster, suggesting that this cluster might represent a well-defined, central grouping of the data. Both Cluster 2 and Cluster 3 each contain only 1 data point. This sparse distribution indicates that, for each clustering algorithm, these clusters do not represent a significant portion of the data. This could suggest that the clusters are either not well-defined or that they are outliers with a high degree of uncertainty in their membership. The standard FCM algorithm takes 10 iterations to converge. This method converges slowly, which might indicate that it requires more time to reach a stable configuration, potentially due to the complexity of the membership assignments in fuzzy clustering. The version of FCM using the Chebyshev distance converges faster, taking only 9 iterations. The Chebyshev distance, which calculates the maximum distance along any dimension, might be more effective in finding the cluster centers and improving the algorithm's convergence speed. The IFCM algorithm, which adds a non-membership function in addition to the membership function, converges in 8 iterations. This suggests that the incorporation of both membership and non-membership parameters may help the algorithm converge more quickly than the standard FCM. The combined approach (IFCM with Chebyshev distance) achieves the fastest convergence, requiring only 6 iterations. The combination of both the Chebyshev distance (which may aid in faster and more robust identification of cluster centers) and the intuitionistic fuzzy approach (which offers a more nuanced clustering mechanism) leads to a significantly faster convergence compared to the other methods. This suggests that this algorithm is more efficient and effective in clustering data, especially when dealing with complex data distributions.

Table 2: Number of features for each cluster.

	Number of Iterations	Cluster 1	Cluster 2	Cluster 3
FCM	10	73	1	1
FCM with Chebyshev	9	73	1	1
IFCM	8	73	1	1
IFCM with Chebyshev	6	73	1	1

While the FCM methods (with and without Chebyshev distance) take more iterations to converge, they may still be suitable for applications where the number of iterations is not a critical factor. However, for faster convergence with similar results, the IFCM with Chebyshev provides the most efficient approach, both in terms of speed (fewer iterations) and adaptability (due to the additional non-membership function). Regardless of the method used, the distribution of data points across the clusters remains the same. This suggests that the clustering methods have similar tendencies when it comes to identifying the core cluster (Cluster 1), even though they differ in their approach to determining the boundaries of the clusters.

Table 3 compares the computational efficiency of three clustering methods: PCA with Fermatean Fuzzy C-Means (FFCM) and Chebyshev, Autoencoders with FFCM and Chebyshev, and IFCM with Chebyshev. The comparison is based on the number of iterations required for convergence and the computational time taken by each method. Among the three methods, PCA-FFCM required the highest number of iterations (10) and the longest computational time (0.02800), making it the least efficient in terms of speed. In contrast, Autoencoders-FFCM had the lowest number of iterations (2) and a shorter computational time (0.0210), indicating faster convergence. However, IFCM-FFCM achieved the best efficiency, requiring only 6 iterations while having the shortest computational time (0.01297), making it the most computationally efficient method. Overall, while Autoencoders-FFCM converges the fastest, its clustering performance was weaker in the previous comparison. IFCM-FFCM emerges as the best choice, as it maintains strong clustering

performance while being the most efficient in terms of computation time and iterations.

Table 3: Comparison between different methods (number of iterations and computational time).

	Number of Iterations	Computational Time
PCA with FFCM and Chebyshev	10	0.02800
Autoencoders with FFCM and Chebyshev	2	0.0210
IFCM with Chebyshev	6	0.01297

Table 4 compares three clustering methods: PCA with FFCM and Chebyshev, Autoencoders with FFCM and Chebyshev, and IFCM with Chebyshev. These methods are evaluated based on five clustering performance metrics. The Silhouette Score, which measures how well-separated the clusters are, is highest for IFCM with Chebyshev (0.9702), indicating the best clustering structure. The Davies-Bouldin Index, where lower values signify better separation between clusters, is also lowest for IFCM (0.0028), confirming minimal overlap between clusters. The Within-Cluster Sum of Squares (WCSS), which measures cluster compactness, is lowest for PCA-FFCM (1.8794), meaning that it forms the most compact clusters. However, in terms of the Calinski-Harabasz Index, where higher values indicate better-defined clusters, IFCM with Chebyshev scores significantly higher (27520.8666), showing the strongest clustering structure. Lastly, the Xie-Beni Index, which assesses both compactness and separation, is lowest for IFCM (7.0248e-08), confirming that it achieves the most optimal clustering. Overall, IFCM with Chebyshev performs the best, as it excels in most of the evaluation metrics, demonstrating well-separated, compact, and well-defined clusters. PCA-FFCM performs best in compactness (WCSS), while Autoencoders-FFCM shows weaker clustering performance compared to the other two methods.

Table 4: Comparison between different methods.

	PCA with FFCM and Chebyshev	**Autoencoders with FFCM and Chebyshev**	**IFCM with Chebyshev**
Silhouette Score [24]	0.9440	0.9188	**0.9702**
Davies-Bouldin Index [38]	0.7576	1.4960	**0.0028**
Within-Cluster Sum of Squares (WCSS) [33]	**1.8794**	29.7129	3.9947
Calinski-Harabasz Index [35]	129.101	27.3406	**27520.8666**
Xie-Beni Index [27]	9.9388	8.2150	**7.0248e-08**

The provided Table 5 compares the performance of four clustering algorithms-FCM, FCM with Chebyshev, IFCM, and IFCM with Chebyshev-across several clustering evaluation metrics. These metrics are used to assess the quality of the clustering solutions, and each algorithm's score in these metrics provides insights into its performance. All four methods have very similar Silhouette Scores around 0.970 (with a slight variation of 0.9703 for FCM and FCM with Chebyshev, and 0.9702 for IFCM and IFCM with Chebyshev), indicating that the clusters are well-separated and compact in all cases. All four algorithms have the same Davies-Bouldin Index value of 0.0028, suggesting that they all perform similarly in terms of cluster separation and compactness. FCM and FCM with Chebyshev have higher WCSS values (around 215.6689 and 216.1086, respectively), indicating that the clusters formed by these methods have higher internal variance compared to

the IFCM methods. IFCM and IFCM with Chebyshev have much lower WCSS values (3.9938 and 3.9947, respectively), indicating that these methods result in more compact clusters. The FCM and FCM with Chebyshev methods have higher Calinski-Harabasz Index values (28110.8762), indicating that these methods achieve better separation between clusters compared to the IFCM methods, which have lower values (27520.8666). FCM and FCM with Chebyshev have values of 1.9263 and 1.9324, respectively, suggesting that these methods result in less optimal cluster quality compared to the IFCM methods. IFCM and IFCM with Chebyshev show extremely low values (7.2590e-08 and 7.0248e-08), indicating excellent cluster compactness and separation. All four methods show very similar performance in terms of the Silhouette Score and Davies-Bouldin Index, suggesting that the clusters are well-separated and compact across all methods. The IFCM algorithms (with and without Chebyshev) show significantly better performance in terms of WCSS and Xie-Beni Index, indicating that these methods generate more compact clusters. The FCM and FCM with Chebyshev methods show better Calinski-Harabasz Index values, indicating they might have better cluster separation compared to the IFCM methods. The IFCM and its variation with Chebyshev distance perform better in terms of compactness and separation, as evidenced by the very low Xie-Beni Index and WCSS values. However, the FCM methods perform slightly better in terms of Calinski-Harabasz Index, which suggests they might provide a better balance between cluster separation and compactness. In conclusion, while the IFCM methods yield the most compact and well-separated clusters overall (as indicated by the low WCSS and Xie-Beni Index), the FCM methods achieve slightly better separation between clusters (as indicated by the Calinski-Harabasz Index). Therefore, the choice of method may depend on whether compactness or separation is prioritized in the given application.

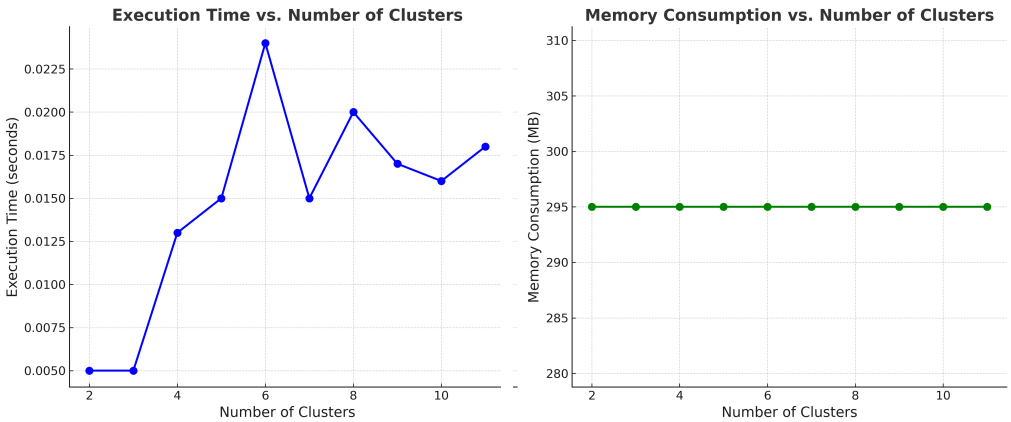
Table 5: Comparison with previous methods.

	FCM	**FCM with Chebyshev**	**IFCM**	**IFCM with Chebyshev**
Silhouette Score [24]	**0.9703**	**0.9703**	0.9702	0.9702
Davies-Bouldin Index [38]	**0.0028**	**0.0028**	**0.0028**	**0.0028**
Within-Cluster Sum of Squares (WCSS) [33]	215.6689	216.1086	**3.9938**	**3.9947**
Calinski-Harabasz Index [35]	**28110.8762**	**28110.8762**	27520.8666	27520.8666
Xie-Beni Index [27]	1.9263	1.9324	**7.2590e-08**	**7.0248e-08**

Figure 8(a) presents the relationship between the execution time (in seconds) and the number of clusters for a clustering algorithm. The execution time increases as the number of clusters increases from 2 to 4. At 4 clusters, the execution time peaks at approximately 0.025 seconds, indicating that the computational complexity or iterations required by the algorithm grow with the number of clusters. At 5 clusters, there is a notable drop in execution time to about 0.01 seconds, suggesting that the algorithm becomes more efficient at this point, possibly due to a better balance in the clustering structure or faster convergence. From 6 to 10 clusters, the execution time fluctuates, with values generally stabilizing around 0.015 to 0.02 seconds. These variations might result from differences in the number of iterations required to converge as the complexity of cluster assignments increases. The execution time peak at 4 clusters suggests that the algorithm's computational demand is highest at this point, possibly due to the increased effort to assign data points accurately to multiple clusters. The drop at 5 clusters might indicate a natural configuration where the algorithm converges more efficiently. The increasing execution time for higher numbers of clusters reflects the typical trade-off between the number of clusters and computational cost. Although the time stabilizes after 5 clusters, the fluctuations suggest that the algorithm handles

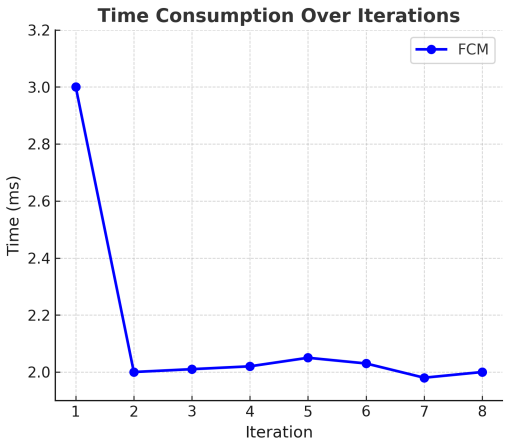
higher cluster numbers with slight efficiency variations. While execution time is a critical factor, it should be considered alongside cluster evaluation metrics like the WCSS, Silhouette Score, and Davies-Bouldin Index. For instance, if 5 clusters provide an efficient execution time with high clustering quality, it might be a practical choice. The execution time analysis demonstrates that the clustering algorithm incurs higher computational costs as the number of clusters increases, with the peak occurring at 4 clusters. However, beyond this, the algorithm shows improved efficiency, stabilizing execution times. This suggests that careful consideration of both execution time and clustering quality metrics is essential for selecting the optimal number of clusters.

Figure 8(b) illustrates the relationship between memory consumption (in MB) and the number of clusters for a clustering algorithm. The memory usage remains stable at approximately 295 MB across all tested cluster numbers (from 2 to 10 clusters). There is no observable variation in memory consumption as the number of clusters increases. Unlike execution time, which typically scales with the number of clusters, memory usage appears to be unaffected. This indicates that the algorithm's memory requirements are independent of the number of clusters in this specific scenario. The stability in memory consumption demonstrates that the algorithm is well-optimized for memory use. It suggests that the internal data structures and computations do not expand significantly with an increasing number of clusters, making the algorithm efficient for memory-constrained environments. The consistent memory footprint implies that the algorithm can handle varying cluster numbers without overburdening system resources, making it suitable for applications involving larger datasets or limited memory systems. Given the constant memory usage, the primary focus for optimizing the algorithm should be on reducing execution time and improving the clustering quality metrics rather than addressing memory consumption. This analysis highlights that the clustering algorithm maintains a fixed memory consumption of around 295 MB regardless of the number of clusters. This stability underscores the algorithm's efficiency in managing memory resources, making it highly scalable and suitable for applications where memory limitations are a concern.

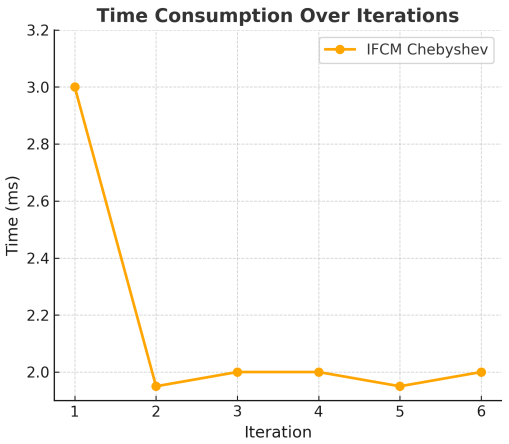


(a) Execution time vs. number of clusters. (b) Memory consumption vs. number of clusters.

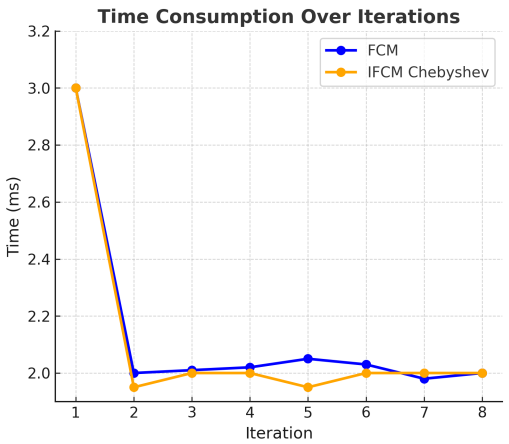
Figure 8: Performance evaluation of the proposed method in terms of (a) execution time and (b) memory consumption with respect to the number of clusters.



(a) Time consumption of FCM over iterations.



(b) Time consumption of IFCM Chebyshev over iterations.



(c) Comparison of time consumption per iteration between FCM and IFCM Chebyshev.

Figure 9: Time consumption trends of FCM and IFCM Chebyshev methods across iterations: (a) FCM over iterations, (b) IFCM Chebyshev over iterations, and (c) comparison per methods.

Figure 9 comprises three subplots, illustrating the time consumption per iteration for different clustering algorithms: IFCM, IFCM with Chebyshev distance, and a comparative analysis of both methods. IFCM time consumption (Figure 9(a)) shows the time consumption per iteration starts high at the first iteration (0.003 seconds) and significantly decreases by the second iteration (0.002 seconds). From the third iteration onward, time consumption stabilizes with minor fluctuations, indicating that most of the computational work occurs in the initial iterations. IFCM with Chebyshev time consumption (Figure 9(b)) shows the first iteration also shows high time consumption (0.003 seconds), followed by a sharp decline to approximately 0.002 seconds in the second iteration. The subsequent iterations (3 to 6) demonstrate stable and consistent time consumption without noticeable fluctuations. Comparison of IFCM and IFCM with Chebyshev (Figure 9(c)) presents both algorithms exhibit similar time consumption trends during the first two iterations, with a sharp decline after the first iteration. Beyond the second iteration, IFCM with Chebyshev appears slightly more stable than IFCM, with fewer fluctuations in time consumption.

Both IFCM and IFCM with Chebyshev demonstrate high computational demand during the initial iteration, likely due to complex initialization or distance calculations. The proposed IFCM with Chebyshev algorithm exhibits a higher computational complexity compared to traditional FCM due to the integration of intuitionistic and the Chebyshev distance metric. This complexity arises from the need to compute multiple membership-related parameters and handle non-linear distance calculations, especially in high-dimensional data environments. Moreover, replacing the traditional Euclidean distance with the Chebyshev distance adds to the computational demand. The Chebyshev metric requires determining the maximum absolute difference across all dimensions, which, while avoiding squaring and square roots, increases the number of comparisons per calculation. In high-dimensional datasets, this leads to a substantial rise in per-iteration computation time, especially during the first few iterations when cluster centers are still unstable and frequent updates occur.

Although the proposed IFCM with Chebyshev increases the computational complexity, the observed decrease in time consumption across subsequent iterations indicates enhanced convergence and reduced processing effort as the cluster centers stabilize. The Chebyshev-based variant exhibits a slightly smoother and more consistent runtime pattern, suggesting greater predictability and stability in computational performance. Both IFCM and its Chebyshev-enhanced counterpart demonstrate a noticeable stabilization in execution time after the second iteration, implying that further iterations contribute minimally to the overall computational cost. This behavior illustrates the efficient computational dynamics of both algorithms. The consistently stable time profile of the Chebyshev variant, in particular, positions it as a favorable choice in applications where computational predictability is essential. These findings underscore the efficiency and reliability of both approaches for iterative clustering tasks. Overall, the enhanced stability, coupled with its superior ability to manage uncertainty and maintain robustness in high-dimensional settings, positions the proposed IFCM with Chebyshev as a more accurate and dependable alternative to traditional FCM, especially for medium-scale datasets where both clustering precision and computational reliability are essential.

5 Comparative Analysis

To evaluate the performance of each proposed method, a sensitivity analysis was conducted under four varying experimental conditions. These conditions include Clustering Accuracy (CA), which measures the percentage of correctly clustered samples compared to the ground truth; Partition Coefficient (PC), which evaluates the sharpness of clustering i.e., whether data points belong strongly to one cluster; Partition Entropy (PE), which measures the fuzziness or uncertainty of clustering, where lower values indicate crisper clustering; and Execution Time (ET), which measures the computational cost or efficiency.

Table 6: Sensitivity analysis.

Method	Avg. CA (%)	Avg. PC (%)	Avg. PE (%)	Avg. ET (s) (%)
FCM	78.5	0.65	0.72	1.12
FCM with Chebyshev	81.2	0.65	0.66	1.25
IFCM	85.7	0.71	0.59	1.45
IFCM with Chebyshev	88.9	0.74	0.52	1.58

Table 6 presents a comprehensive comparison of the four clustering algorithms: FCM, FCM with Chebyshev, IFCM, and IFCM with Chebyshev based on four key performance metrics: CA, PC, PE, and ET. The results demonstrate a clear and consistent trend in which the performance improves progressively from traditional FCM to IFCM-C, particularly in terms of accuracy and clustering clarity. Among all methods, IFCM with Chebyshev achieved the highest clustering accuracy (88.9%), indicating its superior ability to correctly identify and assign data points to their appropriate clusters. This reflects the effectiveness of combining intuitionistic fuzzy logic, which handles uncertainty and hesitation, with the Chebyshev distance metric, which reduces sensitivity to outliers and dimensional variation. IFCM with Chebyshev also recorded the highest Partition Coefficient (0.74) and the lowest Partition Entropy (0.52), which confirms that this method yields crisper, more distinct clusters with less ambiguity in membership assignments.

In contrast, the standard FCM algorithm showed the lowest performance across all metrics, with a clustering accuracy of 78.5%, lower PC (0.65), and higher PE (0.72), suggesting that it struggles more with fuzzy and imprecise data. The use of Chebyshev distance alone in FCM with Chebyshev provided a moderate improvement over standard FCM, particularly in handling noisy and unevenly scaled data. Similarly, IFCM without Chebyshev showed notable improvements due to its ability to capture hesitation through intuitionistic fuzzy sets, resulting in a clustering accuracy of 85.7%, but still lagging behind the full IFCM with Chebyshev approach. However, the increased accuracy and clarity come at the cost of higher computational time. IFCM with Chebyshev recorded the longest execution time (1.58 seconds), followed by IFCM (1.45 s), FCM-C (1.25 s), and FCM (1.12 s). This reflects the additional complexity of computing hesitation degrees and handling maximum-distance calculations in Chebyshev, which adds to the algorithm's processing burden.

In conclusion, Table 6 confirms that IFCM with Chebyshev outperforms all other methods in clustering quality and robustness, particularly in high-dimensional and uncertain data environments. Although it incurs greater computational cost, the trade-off is justified by its superior clustering accuracy and reduced fuzziness [21]. This makes IFCM with Chebyshev a strong candidate for applications where precision, stability, and interpretability are critical, especially in domains like medical diagnosis, fault detection, and decision support systems.

6 Conclusions

The IFCM algorithm, which incorporates the Chebyshev distance metric, serves as an advanced variant of the traditional FCM clustering technique. This method is particularly advantageous for addressing clustering issues when faced with uncertainty and imprecision in datasets. By utilizing IFS, the algorithm introduces an extra layer of flexibility through the handling of hesitation, while the Chebyshev distance metric allows for effective management of multi-dimensional data by focusing on the greatest variation across different dimensions. Our results revealed that the IFCM-C algorithm achieves higher clustering accuracy and integrity, especially in datasets characterized by high levels of ambiguity and multi-dimensionality. The sensitivity analysis further confirmed the consistency of the algorithm's output under different parameter settings and initialization conditions, emphasizing its reliability and generalizability. The integration of these concepts in the IFCM with Chebyshev algorithm enhances its capability to tackle clustering challenges more proficiently. The mechanism of hesitation degrees enables this algorithm to manage uncertainty in data better than conventional fuzzy clustering approaches, making it especially useful when dealing with incomplete or imprecise real-world datasets. IFS provide a flexible framework for representing uncertainty, thus enabling the algorithm to make more informed clustering

choices.

In contrast to the traditional Euclidean distance, which can be affected by outliers, the Chebyshev distance is based on the maximum difference in any single dimension, which helps mitigate the impact of anomalies and variability in the data. This distance metric also effectively manages datasets with varying scales across their dimensions, resulting in a more balanced clustering outcome. The IFCM algorithm adjusts to the underlying data distribution, allowing for more accurate clustering by factoring in both membership and non-membership degrees of data points. Moreover, the combination of Chebyshev distance and intuitionistic fuzzy sets can lead to quicker convergence during iterative optimization processes, which helps in reducing both computational time and effort. The algorithm shows stable performance across diverse datasets, providing consistent clustering results across different executions. However, the algorithm also presents several challenges and limitations. The inclusion of additional degrees (hesitation and non-membership) increases computational complexity, which may become significant with large-scale datasets. Moreover, the performance of IFCM with Chebyshev is highly dependent on careful parameter tuning, such as the fuzzification coefficient and the number of clusters, which may require domain expertise or empirical fine-tuning for optimal results.

Future studies can explore several enhancements to further improve the effectiveness and scalability of the IFCM with Chebyshev algorithm. One potential direction is the integration of adaptive parameter tuning mechanisms using metaheuristic optimization algorithms (e.g., genetic algorithms, particle swarm optimization) to automate the selection of optimal fuzzification parameters and number of clusters. Additionally, extending the algorithm to support online or incremental learning would enable real-time clustering in dynamic and streaming data environments. Another promising area is the hybridization of IFCM with Chebyshev with deep learning-based feature extraction methods to handle unstructured data types such as images and text. Furthermore, applying the algorithm to domain-specific problems such as medical diagnostics, remote sensing, or cybersecurity could provide valuable insights into its practical utility and limitations. Finally, a more extensive benchmarking across diverse, large-scale, and heterogeneous datasets would further validate the algorithm's generalizability and highlight areas for further refinement.

In summary, the IFCM with Chebyshev distance represents a strong and flexible alternative to standard clustering methods, effectively tackling uncertainty and outliers in complex datasets. Although it may require meticulous parameter tuning and can be computationally intensive, its proficiency in managing high-dimensional data makes it a significant asset in data analysis. When compared to other clustering techniques such as K-Means, DBSCAN, and Hierarchical Clustering, the IFCM-C approach offers a more nuanced solution to modern clustering challenges.

Acknowledgement Enormous appreciation and special thanks to the Wedge Group from Cancer Research UK Manchester Centre for their permission to conduct on this study. The authors also would like to thank the Universiti Sultan Zainal Abidin for supporting this research Dana Penyelidikan Universiti 2.0 (UniSZA/2022/DPU2.0/16).

Conflicts of Interest The authors declare no conflict of interest.

References

[1] K. T. Atanassov (1989). More on intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *33*(1), 37–45. https://doi.org/10.1016/0165-0114(89)90215-7.

[2] K. T. Attanassov (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *20*(1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3.

[3] J. C. Bezdek (2013). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media, New York. https://doi.org/10.1007/978-1-4757-0450-1.

[4] A. Bustamam, A. Bachtiar & D. Sarwinda (2019). Selecting features subsets based on support vector machine-recursive features elimination and one dimensional-naïve Bayes classifier using support vector machines for classification of prostate and breast cancer. *Procedia Computer Science*, *157*, 450–458. https://doi.org/10.1016/j.procs.2019.08.238.

[5] H. Chen, S. Das, J. M. Morgan & K. Maharatna (2022). Prediction and classification of ventricular arrhythmia based on phase-space reconstruction and Fuzzy C-means clustering. *Computers in Biology and Medicine*, *142*, Article ID: 105180. https://doi.org/10.1016/j.compbiomed.2021.105180.

[6] T. Y. Chen (2020). New Chebyshev distance measures for Pythagorean fuzzy sets with applications to multiple criteria decision analysis using an extended ELECTRE approach. *Expert Systems with Applications*, *147*, Article ID: 113164. https://doi.org/10.1016/j.eswa.2019.113164.

[7] D. L. Davies & D. W. Bouldin (2009). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

[8] W. Eagles. Subtyping MSc 2021. https://github.com/weaglesBio/SubtypingMSc.

[9] L. M. Elshenawy, C. Chakour & T. A. Mahmoud (2022). Fault detection and diagnosis strategy based on k-nearest neighbors and Fuzzy C-means clustering algorithm for industrial processes. *Journal of the Franklin Institute*, *359*(13), 7115–7139. https://doi.org/10.1016/j.jfranklin.2022.06.022.

[10] Y. Eryoldaş & A. Durmuşoğlu (2022). An efficient parameter tuning method based on the latin Hypercube Hammersley Sampling and Fuzzy C-means clustering methods. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 8307–8322. https://doi.org/10.1016/j.jksuci.2022.08.011.

[11] M. F. Faraloya, S. Shafie, F. M. Siam, R. Mahmud & S. O. Ajadi (2021). Numerical simulation and optimization of radiotherapy cancer treatments using the caputo fractional derivative. *Malaysian Journal of Mathematical Sciences*, *15*(2), 161–187.

[12] M. R. Hassan, M. F. Islam, M. Z. Uddin, G. Ghoshal, M. M. Hassan, S. Huda & G. Fortino (2022). Prostate cancer classification from ultrasound and MRI images using deep learning based explainable Artificial Intelligence. *Future Generation Computer Systems*, *127*, 462–472. https://doi.org/10.1016/j.future.2021.09.030.

[13] J. E. Jackson (2005). *A User's Guide to Principal Components*. John Wiley & Sons, New York. https://doi.org/10.1002/0471725331.

[14] M. Javadian, R. Vaziri, S. Haghzad Klidbary & A. Malekzadeh (2020). Refining membership degrees obtained from Fuzzy C-means by re-fuzzification. *Iranian Journal of Fuzzy Systems*, *17*(4), 85–104. https://doi.org/10.22111/ijfs.2020.5408.

[15] J. O. Kim & C. W. Mueller (1978). *Factor Analysis: Statistical Methods and Practical Issues*. Sage Publications, Beverly Hills, CA.

[16] N. Kumar & H. Kumar (2022). A fuzzy clustering technique for enhancing the convergence performance by using improved Fuzzy C-means and particle swarm optimization algorithms. *Data & Knowledge Engineering*, *140*, Article ID: 102050. https://doi.org/10.1016/j.datak.2022.102050.

[17] N. Kumar, H. Kumar & D. Sharma (2025). Hybrid fuzzy clustering technique to enhance the performance based on a fusion of intuitionistic modified Fuzzy C-means and improved genetic algorithm. *International Journal of Data Science and Analytics*, *20*, 763–786. https://doi.org/10.1007/s41060-023-00474-w.

[18] L. H. Nguyen & S. Holmes (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, *15*(6), Article ID: e1006907. https://doi.org/10.1371/journal.pcbi.1006907.

[19] U. Rusdiana, I. Ernawati, N. Falih & A. Arista (2021). Comparison of distance metrics on Fuzzy C-means algorithm through customer segmentation. In *2021 International Conference on Informatics, Multimedia, Cyber and Information System* (*ICIMCIS*), pp. 307–311. Jakarta, Indonesia. https://doi.org/10.1109/ICIMCIS53775.2021.9699206.

[20] E. H. Ruspini (1969). A new approach to clustering. *Information and Control*, *15*(1), 22–32. https://doi.org/10.1016/S0019-9958(69)90591-9.

[21] M. S. Saidin, L. S. Lee, M. R. A. Bakar & M. Z. Ahmad (2022). A new divergence measure based on fuzzy TOPSIS for solving staff performance appraisal. *Malaysian Journal of Mathematical Sciences*, *16*(3). https://doi.org/10.47836/mjms.16.3.14.

[22] M. E. Salman, G. Ç. Çakar, J. Azimjonov, M. Kösem & İ. H. Cedimoğlu (2022). Automated prostate cancer grading and diagnosis system using deep learning-based Yolo object detection algorithm. *Expert Systems with Applications*, *201*, Article ID: 117148. https://doi.org/10.1016/j.eswa.2022.117148.

[23] J. Serey, M. Alfaro, G. Fuertes, M. Vargas, C. Duran, R. Ternero, R. Rivera & J. Sabattin (2023). Pattern recognition and deep learning technologies, enablers of industry 4.0, and their role in engineering research. *Symmetry*, *15*(2), Article ID: 535. https://doi.org/10.3390/sym15020535.

[24] K. R. Shahapure & C. Nicholas (2020). Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics* (*DSAA*), pp. 747–748. IEEE, Birmingham, UK. https://doi.org/10.1109/dsaa49011.2020.00096.

[25] F. J. Shaikh & D. S. Rao (2022). Prediction of cancer disease using machine learning approach. *Materials Today: Proceedings*, *50*(1), 40–47. https://doi.org/10.1016/j.matpr.2021.03.625.

[26] J. Shi, Y. Ye, D. Zhu, L. Su, Y. Huang & J. Huang (2021). Comparative analysis of pulmonary nodules segmentation using multiscale residual U-Net and Fuzzy C-means clustering. *Computer Methods and Programs in Biomedicine*, *209*, Article ID: 106332. https://doi.org/10.1016/j.cmpb.2021.106332.

[27] M. Singh, R. Bhattacharjee, N. Sharma & A. Verma (2017). An improved xie-beni index for cluster validity measure. In *2017 Fourth International Conference on Image Information Processing (ICIIP)*, pp. 1–5. IEEE, Shimla, India. https://doi.org/10.1109/ICIIP.2017.8313691.

[28] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal & F. Bray (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660.

[29] A. A. Supianto, N. Sa'diyah, C. Dewi, R. I. Rokhmawati, S. A. Wicaksono, H. M. Az-Zahra, S. H. Wijoyo, Y. Hayashi & T. Hirashima (2020). Improvements of Fuzzy C-means clustering performance using particle swarm optimization on student grouping based on learning activity in a digital learning media. In *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*, pp. 239–243. Association for Computing Machinery, New York. https://doi.org/10.1145/3427423.3427449.

[30] S. Surono & R. D. A. Putri (2021). Optimization of Fuzzy C-means clustering algorithm with combination of Minkowski and Chebyshev distance using principal component analysis. *International Journal of Fuzzy Systems*, *23*(1), 139–144. https://doi.org/10.1007/s40815-020-00997-5.

[31] E. Szmidt & J. Kacprzyk (2000). Distances between intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, *114*(3), 505–518. https://doi.org/10.1016/S0165-0114(98)00244-9.

[32] K. Tabianan, S. Velu & V. Ravi (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, *14*(12), Article ID: 7243. https://doi.org/10.3390/su14127243.

[33] E. Umargono, J. E. Suseno & S. K. V. Gunawan (2020). K-means clustering optimization using the Elbow method and early centroid determination based on mean and median formula. In *The 2nd International Seminar on Science and Technology (ISSTEC 2019)*, pp. 121–129. Atlantis Press, Dordrecht, Netherlands. https://doi.org/10.2991/assehr.k.201010.019.

[34] B. Wang, S. Hu, Y. Teng, J. Chen, H. Wang, Y. Xu, K. Wang, J. Xu, Y. Cheng & X. Gao (2024). Current advance of nanotechnology in diagnosis and treatment for malignant tumors. *Signal Transduction and Targeted Therapy*, *9*(1), Article ID: 200. https://doi.org/10.1038/s41392-024-01889-y.

[35] X. Wang & Y. Xu (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In *IOP Conference Series: Materials Science and Engineering*, volume 569 pp. Article ID: 052024. IOP Publishing, Bristol, England. https://doi.org/10.1088/1757-899X/569/5/052024.

[36] W. E. Wright (1973). A formalization of cluster analysis. *Pattern Recognition*, *5*(3), 273–282. https://doi.org/10.1016/0031-3203(73)90048-4.

[37] C. Wu & X. Zhang (2022). A self-learning iterative weighted possibilistic Fuzzy C-means clustering via adaptive fusion. *Expert Systems with Applications*, *209*, Article ID: 118280. https://doi.org/10.1016/j.eswa.2022.118280.

[38] J. Xiao, J. Lu & X. Li (2017). Davies Bouldin index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327–1338. https://doi.org/10.3233/IDA-163129.

[39] Z. Xu & J. Wu (2010). Intuitionistic Fuzzy C-means clustering algorithms. *Journal of Systems Engineering and Electronics*, *21*(4), 580–590. https://doi.org/10.3969/j.issn.1004-4132.2010.04.009.

[40] L. A. Zadeh (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X.

[41] N. Zamri, N. A. A. Bakar, A. Z. Abd Aziz, E. N. Madi, R. A. Ramli, S. M M & C. S. Koon (2024). Development of Fuzzy C-Means with Fuzzy Chebyshev for genomic clustering solutions addressing cancer issues. *Procedia Computer Science*, *237*, 937–944. https://doi.org/10.1016/j.procs.2024.05.182.