International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

SCIE_EOC -A Sinusoidal Chaotic and Information Entropy based Elephant-Herding Optimization for Clustering the Large Datasets to Improve the Data Search Capability

Dr Pradeep Kumar Atulker

Assistant Professor, Department of Computer Science and Application, Govt. College Khimlasa, Sagar, (M.P.), India

Dr Rajendra Gupta

Associate Professor, Department of CS & IT, Rabindranath Tagore University, Raisen (M.P.), India,

Dr Ankur Khare

Assistant Professor, Department of CS & IT, Rabindranath Tagore University, Raisen (M.P.), India,

Abstract - The data clustering is extensively employed for a vast diversity of applications in many sectors, like learning, curative practices, innovation, and enetrprises in diverse decentralized scenarios. The substantial diverse data are managed and scrutinized using multiple categorization techniques to optimize the excellence of crucial information exchange across decentralized scenarios. In this paper, a Sinusoidal Chaotic and Information Entropy based Elephant-Herding Optimization (SCIE_EOC) scheme is developed and applied to classify the significant volume of diverse datasets, aiming to optimize the data search capability. The primary focus of SCIE_EOC is to attain proficient and precise dispersal of data values in dataset by incorporating the principles of information entropy and employing sinusoidal chaotic for population formation. The incorporation of sinusoidal chaotic and information entropy are intended to boost the assessment and utilization abilities of search agents in optimization scheme and to achieve optimal and exact assortment of cluster center and members. The execution of SCIE_EOC scheme utilizing MATLAB 2022a software on six extensive datasets reveals the remarkable efficacy of SCIE_EOC scheme in comparison to earlier schemes like K-Means, PSO, GWO, and EO as evidenced by factors including Root Mean Square Error, F-Measure, Standard Deviation, Purity Index, Accuracy, Intra-Cluster Distance, Time Complexity and Statistical Exploration.

Keywords: Clustering, Elephant-Herding Optimization, F-Measure, Information Entropy, Purity Index, Sinusoidal Chaotic, Statistical Exploration.

1. INTRODUCTION

The researchers discuss the handling and exploration of big data [1, 2] in many applications, leveraging aspects like storage, performance, and database administration. Pre-processing is utilized to reduce data degradation and boost data accuracy and productivity. A dominant dispersed architecture is employed, enabling fast and economical data examination for both serial and parallel communication [3, 4]. A number of population based optimization scehems [22], like Grey Wolf Optimization (GWO) [5], Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), are used to optimally and accurately examine the vast volume of data in several scenarios [6, 7, 8]. The ornamental cognitive and deep learning using neural networks are applied in big data exploration; in which, case related perceptive is employed for query handling, aiming to moderate query retrieving rate and boost retrieving time. Health information is examined to depict illness with qualitative and quanitative justifications, guranteeing most appropriate medical care [10, 11].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Convex optimization is emphasized as one of the top techniques for large-scale data examination in dissiminated environment, particularly for diverse sources. The convex signal handling techniques is leveraged with the tiniest square scheme for sampling and estimation of millions of data. The proximal operator, unpredictability and disseminated estimation are essential aspects in the optimization of data examination [13, 26]. The deep learning is helped in the execution of dispersed clustering on diverse data utilizing an auto-encoder. The auto-encoder autonomously determines the count of centers and assigns members of clusters with the help of deep learning scheme. The outcomes are examined with the renowned clustering scheme, K-Means [15, 27].

Additionally, a amalgamation of Convolutional Neural Network (CNN) and machine learning schemes is employed to enhance transparency and strictness in disease category picking and handling. The parallel computation, including the auto-encoder, is employed to choose the best grouping strategies for medical health administration and perception established learning around sicknesses [16]. The decision making regarding sickness is executed with the help of artificial intelligence, verified and validated on medical database. Various forms of cancers are examined utilizing support vector machine to ensure appropriate and exact disease handling [17]. The machine learning is additionally utilized with the help of tensor and central operators to efficiently handle and validate the datasets; where, the data validation is carried out through deep learning, applied in diverse platforms like Yahoo and dynamic applications [19, 20].

Grouping, akin to clustering, involves classify numerous datsets into categories for diverse aims. In the medical context, a decision support scheme is employed to group clinical data and forecast diseases [21, 29]. For categorization of medical information, a CNN is utilized with an innovative loss function. The multi-pool system is utilized to confinement the spatial information of primary features in healthcare information, assisting in assessing the severity of disease and deciding suitable therapies [23].

The above literature shows the use of various clustering schemes and optimization schemes like GWO [5], GA and PSO, and K-Means [15, 27] on different datasets to improve data extraction and data searching capability. While some clustering techniques show promising results on particular datasets, their efficiency may vary on other datasets due to the no-free-lunch theorem, indicating that no single clustering scheme is universally optimal for all problems. The rate of convergence is also a primary consideration of optimization methods.

To address these challenges, a new approach called a Sinusoidal Chaotic and Information Entropy based Elephant-Herding Optimization for Clustering (SCIE_EOC) scheme is introduced. This scheme is applied to cluster extensive datasets and is an upgraded version of the Elephant-Herding Optimization (EO) scheme. Unlike traditional qualitative analyses of prior works, the proposed SCIE_EOC scheme focuses on a quantitative consideration of clustering schemes for large datasets.

The main contribution of SCIE_EOC implementation is as follows:

1. The significant objective of SCIE_EOC is to attain an effective dissemination of data component values across diverse resources in a distributed scenario.

2. The data dissemination is accomplished by leveraging sinusoidal chaotic [24, 25] for population generation and information entropy for data values distribution.

3. The ultimate goal is to optimize the distribution of data elements, ensuring efficient utilization of heterogeneous resources within the distributed system.

4. The implementation of SCIE_EOC scheme utilizing MATLAB 2022a software on six extensive datasets as evidenced by factors including Root Mean Square Error, F-Measure, Standard Deviation, Purity Index, Accuracy, Intra-Cluster Distance, Time Complexity and Statistical Exploration in comparison to earlier schemes like K-Means, PSO, GWO, and EO.

Part 1 provides an introduction, while Part 2 presents a literature survey on data clustering and classification schemes, and big data examination. This section explores various parameters related to these fields. In Part 3, the Elephant-Herding Optimization (EO) scheme is discussed, laying the foundation for the proposed SCIE_EOC

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

approach, which is explained in Part 4. Part 4 includes complete details such as flowcharts, and algorithms related to the SCIE_EOC approach. Part 5 focuses on datasets used in the experiments, performance metrics, and the results obtained from applying the SCIE_EOC approach. It sheds light on how well the proposed method performs in practice. Lastly, Part 6 comprises the conclusions drawn from the study, summarizing the key findings and implications of the research.

2. LITERATURE REVIEW

The Decisiion support systems (DSS) [28] discusses the expansion of machinery to manage numerous complex data retrieving from distributed sources. It introduces the concepts of Big Data [32, 34], which are high-tech multidimensional information controlling structures that aid investors in using up-to-date data-driven schemes for problem resolving. The ubiquity of Big Data in nutrition protection is emphasized, as data in the nutrition supply chain is dispersed and diverse in setup, measure, and terrestrial source. Overall, the examination and conversation delivered purpose to encourage the use of Big Data and DSS in nutrition protection, ultimately improving risk impost and ensuring nutrition protection in the context of climate variation.

The clustering [12] is as a widely utilized strategy for grouping data into comparable clusters. However, most existing systems for clustering steadly describe clusters with instances values, and concentration, lacking perfect semantic sense. To address this issue, a new Evolutionary Clustering Algorithm (ECA) is developed that integrates communal class position, population based schemes, Levy flight optimization, and K-Means scheme. The evaluation involves using 32 diverse datasets, assessing their efficiencies using centre and outside clustering actions. In conclusion, the proposed ECA algorithm presents a significant improvement in clustering heterogeneous and multiple-featured datasets with better cluster identification and increased robustness across various dataset characteristics.

The study [30] focuses on association rule mining strategies and their application in college student quality valuation schemes. These strategies are used to discover probable relations between diverse abilities of college students based on their existence and knowledge data. This helps teachers identify individual students' powers and weaknesses and enables personalized teaching. The objective is to address issues related to associative rule removal schemes and explore the influence of implementing such schemes in college student excellence valuation. The experiment confirms the strength and achievability of the scheme for college student excellence valuation, supporting its practical application in real-world scenarios. Overall, the paper highlights the benefits of using association rule mining strategies in college student excellence valuation systems, paving the way for more effective and personalized teaching strategies.

The paper introduces a novel heuristic approach for data clustering based on the Moth Flame Optimizer (MFO) [9], a new metaheuristic inspired by the intelligence of moths in nature. While the k-means scheme is widely recognized for clustering, its procedure is highly reliant on primary cluster centers and may get trapped in local optima because of weak search proficiency. The MFO is proposed as a potential solution to manage complex problems, and its act has been establish acceptable in numerous researches. To evaluate the MFO effectiveness, experiments are accompanied employing UCI datasets.

The paper [14] discusses the cumulative proficiency to gather a enormous volume of diverse data, with a particular focus on time-series data, which holds a significant amount of untapped information. Previous data mining schemes often have limitations when examining time-series, particularly when dealing with multi-dimensional time-series that need to be examined together to remove meaningful information. In response, the paper implements a different clustering scheme called K-MDTSC (K-Multi-Dimensional Time-Series Clustering), exactly intended to address multi-dimensional time-series clustering. The outcomes illustrate that K-MDTSC produce decent clustering outputs, particularly when dealing with more complicated synthetic datasets. Overall, the paper highlights the significance of multi-dimensional time-series clustering and demonstrates the effectiveness of the proposed K-MDTSC algorithm in both synthetic and real-world scenarios. The findings suggest that K-MDTSC can be a valuable tool in extracting knowledge from complex multi-dimensional time-series data, particularly in industrial applications such as predictive maintenance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The PSO [31] is a commonly used method to calculate the universal finest of multivariable objevtive functions. In this study, a new PSO scheme is established to exactly calculate the finest value of multidimensional objective functions. The PSO repetition procedure contains of numerous huge repetition steps, every including two phases. In 1st phase, a development practice is employed to successfully discover the multidimensional adjustable area. The 2nd phase employs the traditional PSO scheme to calculate the universal finest value of the function. Overall, the paper presents a new PSO algorithm designed for multidimensional function optimization, demonstrating its effectiveness and superior accuracy through rigorous testing and analysis. This algorithm has the potential to be valuable in various scientific and engineering applications.

The paper [33] focuses on the operational description of clusters to understand their magnitude and composition related features. As tentative schemes only may not offer a whole image of cluster organizations, autonomous theoretical examinations are desired to attain a comprehensive explanation of the symmetrical procedure and features of the clusters. To determine numerous minima and eventually realize the universal minima for the clusters, potential energy surfaces (PES) are discovered. Various optimization schemes are developed and utilized for resolving geometrical isomers in clean essential groups, like Genetic Algorithm (GA). In this article, the optimization schemes, evaluation methods and precision of outcomes achieved by employing the strategies for various cluster type is to be examined.

The paper [18] addresses the serious concern of the Dimensionality Curse, which delays progress in various fields, including bioinformatics. To overcome this challenge, the authors propose a corporation solution that combines scrambling based dimensionality decline scheme with PSO based evaluation methods. The paper focuses on feature assortment for cancer forecast using a complex Spark Distributed scheme called Spark Distributed PSO (SDPSO). The proposed approach integrates Binary PSO (BPSO) with PSO scheme to achieve efficient and effective feature selection. Overall, the proposed SDPSO approach proves to be highly efficient and accurate in feature selection for cancer forecast, showcasing its potential to address the Dimensionality Curse and improve performance in bioinformatics applications.

3. THE ELEPHANT-HERDING OPTIMIZATION (EO) SCHEME

The widespread optimization problems are deteched by the EO, which is a bio-inspired method copied of the nature of elephant herding. There are three general conceptions in EO; (a) Clans are integrated by uniting numerous predeterminated elephants, and the elephant`s population is attained by uniting multiple clans; (b) the movement of different predeterminated male elephants from their clans and their strength of movement from the optimal elephant swarm at every establishment are required; (c) in the clan, the elephants are similarly stayed with the support of a matriarch.

3.1. Clan Updation Scheme

The location of succeeding elephants in clan L_n is obstructed through matriarch. The location of α^{th} elephant is reformed in clan L_n using eq (1).

$$X_{next,L_n,\alpha} = X_{L_n,\alpha} + \beta \times \left(X_{optimal,L_n} - X_{L_n,\alpha}\right) \times rm \quad (1)$$

Where,

$X_{next,L_n,\alpha}$ & $X_{L_n,\alpha}$ = The new and current locations of α^{th} elephant in clan L_n respectively.

β = A magnitude term attaning the influence of matriarch L_n on $X_{Ln,\alpha}$. $\{\beta \in [0,1]\}$

$X_{optimal,L_n}$ = The matriarch (an elephant having highest fitness value in clan L_n).

rm = Random number. $\{rm \in [0,1]\}$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

The location of α^{th} elephant having highest fitness value is reformed in every clan using eq. (2). (Here, reformation is not performed using eq. (1), i.e., $X_{L_n,\alpha} = X_{optimal,L_n}$)

$$X_{next,L_n,\alpha} = \delta \times X_{mid,L_n} \quad (2)$$

Where,

δ = A term attaining the rein of X_{mid,L_n} on $X_{next,L_n,\alpha}$. $\{\delta \in [0,1]\}$

X_{mid,L_n}= The center value of Clan L_n.

The location of α^{th} elephant (eq. (3)) is reformed by adding η^{th} dimension in eq. (2). $\{1 \leq \eta \leq D, \quad D = Overall\ Dimensions\}$

$$X_{mid,L_n,\eta} = \frac{1}{T_{L_n}} \times \sum_{\alpha=1}^{T_{L_n}} X_{L_n,\alpha,\eta} \quad (3)$$

Where,

T_{L_n}= Overall elephants in clan L_n. (T_{cn} = Overall clans in elephant`s population)

$X_{mid,L_n,\eta}$ = The center value of elephant location in clan L_n with η^{th} dimension.

$X_{L_n,\alpha,\eta}$= The location of α^{th} elephant in clan L_n (η^{th} dimension).

3.2. Separating Scheme

Onward, enhancing the investigation strength of EO scheme, comply that the separating scheme will identify the elephants devising worst fitness value at each iteration as expressed in eq. (4).

$$X_{wr,L_n} = X_{lw} + (X_{et} - X_{lw} + 1) \times rd \quad (4)$$

Where,

X_a & X_i= Extreme and Lowest limit of elephant location.

rd = Notional and reliable distribution. $\{rd \in [0,1]\}$

X_{wr,L_n}= Elephant having worst fitness value in the clan L_n.

4. SCIE_EOC - A SINUSOIDAL CHAOTIC AND INFORMATION ENTROPY BASED ELEPHANT-HERDING OPTIMIZATION FOR CLUSTERING

The EO scheme is modified with the help of sinusoidal chaotic and information entropy. The effectual consequences for numeric optimization are attained by the EO scheme, which is right fit for disquisitioning and deployment in a massive searching region. However, the data clustering scheme is genuinely incongruent from numeric optimization. At present, the adjustment of numerous essential features is performed, and the application of a sinusoidal chaotic and information entropy for data clustering is carried out using a SCIE_EOC.

4.1. Component Information Entropy

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Within the data clustering tactic, the dispersion of data points occurs in a multi-dimensional region. The level of ambiguity of a probabilistic variable is computed using information entropy; subsequently, this is employed for measurements of components to determine the distribution. The entropy (P) is assessed to separate the values of components by approximating each value to its nearest whole number using eq. (5).

$$P_q = -\sum_{u=m_q}^{M_q} G_u \log(G_u) \qquad (5)$$

Where,

P_q = q^{th} component entropy (q = 1,2,……..,T)

T = Total components in dataset

M_q & m_q = q^{th} component maximum and minimum value when categorized.

G_u = u^{th} component numerical percentage value.

The maximum entropy is computed for each component in dataset according to eq. (6) during the clan updation scheme of EO population.

$$MAX(P_q) = -\log\left(\frac{1}{M_q - m_q + 1}\right) \qquad (6)$$

Where,

$MAX(P_q)$ = The q^{th} component maximum entropy in dataset.

Finally, eq. (7) is used to compute the normalized entropy for each component.

$$NORM(P_q) = \frac{P_q}{MAX(P_q)} \qquad (7)$$

Where,

$NORM(P_q)$ = The q^{th} component normalized entropy in dataset.

Eq. (7) is iteratively applied to complete data components in dataset, resulting in the generation of a set of normalized entropy (NORM(P)).

NORM (P) = (NORM(P_1), NORM(P_2), ……..,NORM(P_T))

4.2. Information Entropy based Clan Updation in EO Scheme

In this context, two mechanisms are incorporated in clan updation to enhance the efficiency of EO scheme. The initial one is original clan updation scheme of EO elucidated in the preceding section. The second mechanism (Algorithm 1) is the revised version of the original clan updation, where O individuals are randomly selected from the current population ($1 \leq O \leq O^P$), and the best individual is designated as the reference individual ($X_{ref,L_n,\alpha}$). The clan direction is steered by $X_{ref,L_n,\alpha}$, while information entropy is employed to establish an equivalence between population diversity and convergence rate. The increased information entropy of a component denotes the heightened indeterminacy, which adversely affects the rate of convergence. Therefore, the velocity is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

improved for a component by relocating it to the position on the basis of the reference individual, universally with the highest probability in contrast to the component with the least information entropy.

Algorithm 1: Information Entropy based Clan Updation in EO Scheme

Algorithm	**Number of Operations**
START	
FOR y = 1 to O^P **DO**	(O^P+1)
Pick the optimal choice from O random selections as $X_{ref,L_n,\alpha}$	O^P
FOR z = 1 to T_i **DO**	$O^P*(T_i+1)$
IF arbitrary < NORM (P_q) then	O^P*T_i
$X_{next,L_n,\alpha} = X_{ref,L_n,\alpha}$	
ELSE	
Revise $X_{L_n,\alpha}$ and generate $X_{next,L_n,\alpha}$ (eq. (1))	
END IF	
END FOR	
END FOR	
STOP	

4.3. Sinusoidal Chaotic based Selection of Population of EO Scheme

A upgraded EO is devised using a sinusoidal chaotic to counter imprudent convergence calamity. The inclusion of a sinusoidal chaotic in EO scheme is intended to counter the issue of local optima through the utilization of pseudorandom numbers. The sinusoidal chaotic is assessed to acquire the three parameters (τ, j_1, j_2) utilizing eq. (8) to eq. (11).

$$\tau_{m+1}^{next} = \sigma(\tau_m^{next})^2 \sin(\pi\tau_m^{next}) \quad (8)$$

$$j_{1,m+1}^{next} = \sigma\left(j_{1,m}^{next}\right)^2 \sin\left(\pi j_{1,m}^{next}\right) \quad (9)$$

$$j_{2,m+1}^{next} = \sigma\left(j_{2,m}^{next}\right)^2 \sin\left(\pi j_{2,m}^{next}\right) \quad (10)$$

$$\left(\tau_0^{next}, j_{1,0}^{next}, j_{2,0}^{next}\right) \in [0,1], \sigma \in (0,4] \quad (11)$$

Where,

m = Overall iterations

The primary values of EO population (O^p) are determined by assessing the sinusoidal chaotic (eq. (8) to eq. (11)), aiming to enhance the effectiveness of EO scheme through suitable utilization of a vast solution space.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

4.4. The Entire SCIE_EOC Scheme for Data Clustering

The amalgamation of information entropy-driven clan updation and sinusoidal chaotic-driven selection of population is employed to achieve an optimal solution for data clustering in SCIE_EOC (Fig. 1) (Algorithm 2). To begin with, eq. (7) is iteratively applied to complete data components in dataset, ensuing in the production of a set of normalized entropy (NORM(P)). The sinusoidal chaotic (eq. (8) to eq. (11)) is employed to prepare the population (O^P) of the EO scheme, where each individual is represented as a vector comprising $T_i = T \times R$, (T_i = Individual dimension, T = Total components in dataset, R = Total cluster centers). The R cluster centers positions are set in a vector, with each cluster center being linked to its respective set of T attributes. The initial specific vector values are randomly and uniformly created within the range of lower and upper component values present in dataset for a iteration threshold (ITR). Next, SCIE_EOC is employed to compute the fitness values of all individuals using eq. (1) to eq. (4).

Algorithm 2: The Entire SCIE_EOC Scheme

Algorithm	**Number of Operations**
START	
Allocate iteration counter w =1 and iteration threshold (ITR)	1
Compute the entropy for each component to generate normalized entropy (NORM(P))	1
Prepare the population O^P of EO by employing sinusoidal chaotic (eq. (8) to eq. (11))	1
Determine fitness for every elephant	O^P
WHILE (w < ITR) **DO**	ITR+1
Do Sorting to overall elephants through their fitness	ITR
FOR L_n = 1 to T_{cn} **DO**	ITR*(T_{cn}+1)
FOR α = 1 to T_{L_n} **DO**	ITR*T_{cn}*(T_{Ln}+1)
Do information entropy based clan updation (Algorithm 1)	2*O^P*T_i+3*O^P+1
Revise $X_{L_n,\alpha}$ and generate $X_{next,L_n,\alpha}$ (eq. (1))	ITR*T_{cn}*T_{Ln}
IF $X_{L_n,\alpha} = X_{optimal,L_n}$ **THEN**	ITR*T_{cn}*T_{Ln}
Revise $X_{L_n,\alpha}$ and generate $X_{next,L_n,\alpha}$ (eq. (2))	ITR*T_{cn}*T_{Ln}
END IF	
END FOR	
END FOR	
FOR L_n = 1 to T_{cn} **DO**	ITR*(T_{cn}+1)
Adjust the L_n clan elephant with worst fitness (eq. (4))	ITR*T_{cn}

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

END FOR

Compute population by freshly revised positions ITR

w = w+1; ITR

END WHILE

Return optimal elephant with maximum fitness 1

STOP

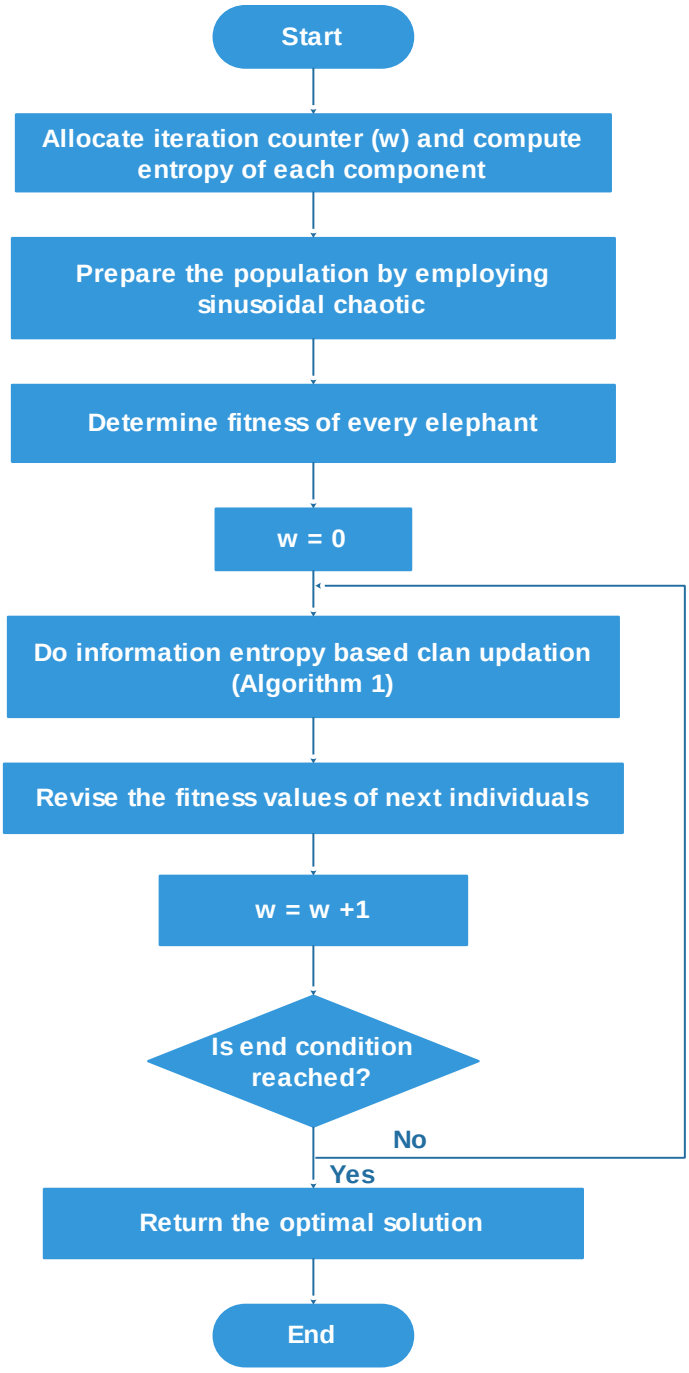

Fig. 1: The flowchart of SCIE_EOC Scheme

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

5. RESULTS AND ANALYSIS

This segment presents a concize overview of the six extensive datasets (Table 1) and performance metrics for the SCIE_EOC scheme. All the schemes are implemented within the MATLAB 2022a tool running on a windows 10 operating system and validated on six comprehensive datasets. The results for the SCIE_EOC scheme are acquired through 500 iterations, considering metrics like Root Mean Square Error, F-Measure, Standard Deviation, Purity Index, Accuracy, Intra-Cluster Distance, Time Complexity and Statistical Exploration. Comparative analysis is conducted against earlier scheme like K-Means, PSO, GWO, and EO over 25 distinct runs.

5.1. Datasets

The SCIE_EOC scheme is executed on six substantial, separate datasets obtained from UCI repository. The datsets are HCV, Blood Transfusion, Thoraic Surgery, HTRU, Diabetic Retinopathy and Cancer (Table 1).

Table 1: Datasets

Sr. No.	Datasets	Total Components/ Instances	Total Features/Attributes	Total Classes/Clusters
1	HCV	1385	29	4
2	Blood Transfusion	748	4	2
3	Thoraic Surgery	470	17	2
4	HTRU	12330	9	2
5	Diabetic Retinopathy	1151	18	2
6	Cancer	683	9	2

5.2. Performance Metrics

The SCIE_EOC scheme is assessed considering metrics like Root Mean Square Error, F-Measure, Standard Deviation, Purity Index, Accuracy, Intra-Cluster Distance, Time Complexity and Statistical Exploration.

5.2.1. Standard Deviation

The fixed data grouping within the mean value domain is attained using a geometric characteristic reffered to as Standard Deviation (V). The finest grouping is achieved by employing the minimum value of V (eq. (12)).

$$V = \sqrt{\frac{\sum(A - \bar{A})}{|T|}} \qquad (12)$$

Here,

$|T|$= Dataset Length

A = Dataset Components Values

$\bar{A}$ = Mean Value of Components in Dataset

Table 2: Standard Deviation

Datasets	K-Means	PSO	GWO	EO	SCIE_EOC

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

HCV	0.13534	0.05263	0.03253	0.03036	0.00892
Blood Transfusion	0.00463	0.18393	0.14342	0.17386	0.00543
Thoraic Surgery	0.00463	0.18364	0.00327	0.00183	0.00352
HTRU	0.27831	0.19373	0.16384	0.14253	0.11854
Diabetic Retinopathy	0.23748	0.11623	0.06274	0.04854	0.00546
Cancer	0.04253	0.12643	0.11837	0.08743	0.00846

According to Table 2 and Fig. 2, the SCIE_EOC determined the least value of V for the six comprehensive dataset. In Standard Deviation (V), the SCIE_EOC achieves 13%, 24%, 38% and 54% higher performance than EO, GWO, PSO and K-Means respectively, across all six datasets.

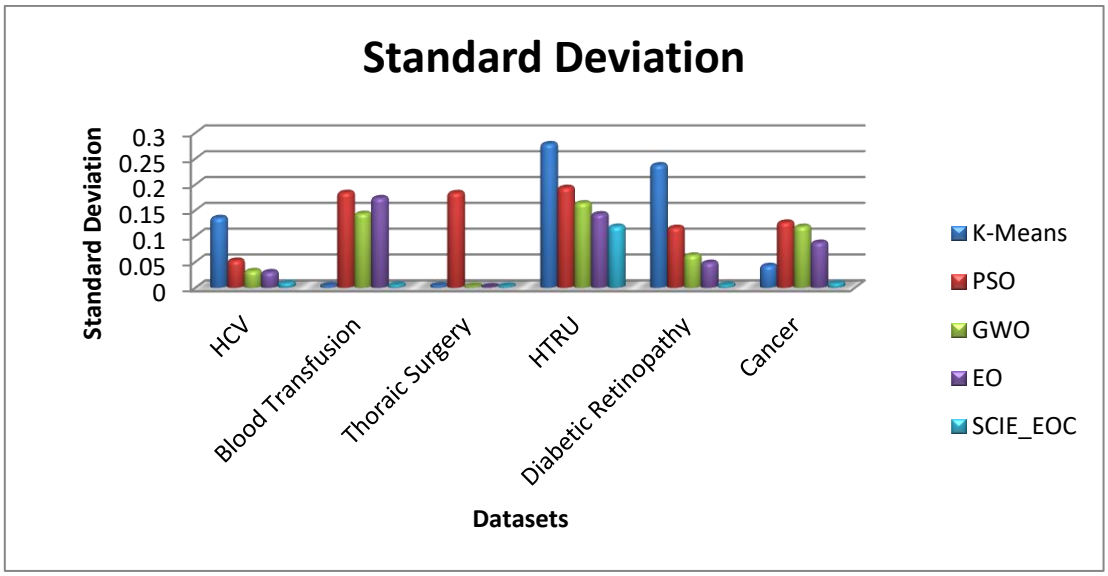

Fig. 2: Standard Deviation

5.2.2. Root Mean Square Error

The value of Root Mean Square Error (E) quantifies the discrepancy between predicted and computed values. The finest grouping is achieved by employing the minimum value of E (eq. (13)).

$$E = \sqrt{\frac{1}{|T|}\sum_{b=1}^{|T|}\left(A_b - \hat{A}_b\right)^2} \qquad (13)$$

Here,

A_b& $\hat{A}_b$ = bth component computed nad predicted value in dataset.

Table 3: Root Mean Square Error

Datasets	K-Means	PSO	GWO	EO	SCIE_EOC
HCV	0.473	0.273	0.098	0.036	0.016
Blood Transfusion	0.427	0.275	0.294	0.193	0.074
Thoraic Surgery	0.538	0.285	0.189	0.154	0.034

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

HTRU	0.395	0.296	0.353	0.173	0.045
Diabetic Retinopathy	0.486	0.376	0.186	0.039	0.032
Cancer	0.546	0.365	0.276	0.143	0.017

The data in Table 3 and Fig. 3 reveal that the SCIE_EOC yielded the least value of E for the six comprehensive dataset. According to Root Mean Square Error (E), the SCIE_EOC surpasses EO, GWO, PSO and K-Means by 17%, 59%, 63% and 84% respectively, across all six datasets.

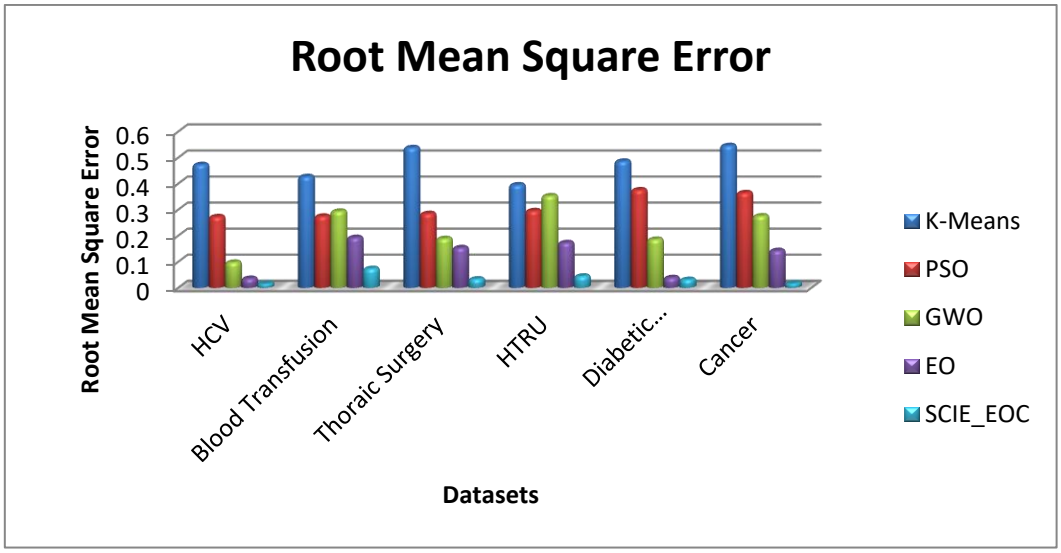

Fig. 3: Root Mean Square Error

5.2.3. Intra-cluster Distance

To begin with, the distances between data components within a cluster are computed. Afterward, the intra-cluster distance is determined as the mean of these distances. The finest grouping is achieved by retaining the minimum value of intra-cluster distance. By computing the mean of distances from the group center to entire data componentes within a group, the average distance is obtained, and this process is repeated for each group. Ultimately, the mean intra-cluster distance is computed by aggregating the mean distances from all groups.

Table 4: Average Rank (Intra-Cluster Distance)

Dataset	K-Means	PSO	GWO	EO	SCIE_EOC
HCV	362.46 (5)	283.24 (4)	212.74 (3)	184.34 (2)	142.37 (1)
Blood Transfusion	4.7352 (4)	5.6345 (5)	4.3654 (3)	2.8643(2)	2.6453 (1)
Thoraic Surgery	2152.8 (5)	1432.2 (3)	1927.8 (4)	1214.2 (2)	1164.3 (1)
HTRU	74.372 (3)	82.383 (4)	94.287 (5)	64.382 (2)	53.251 (1)
Diabetic Retinopathy	8.2872 (5)	6.3423 (3)	7.2537 (4)	5.3627 (2)	3.2836 (1)
Cancer	36.382 (5)	27.386 (4)	19.283 (3)	15.374 (2)	9.8353 (1)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Average Rank (I_φ)	4.5	3.83	3.67	2	1

According to Table 4 and Fig. 4, the SCIE_EOC determined the least value of intra-cluster distance for the six comprehensive dataset. In intra-cluster distance, the SCIE_EOC achieves 18%, 32%, 43% and 51% higher performance than EO, GWO, PSO and K-Means respectively, across all six datasets. The schemes are ranked based on the average of their intra-cluster distances, ranging from the minimum to maximum (from 1 to 4.5).

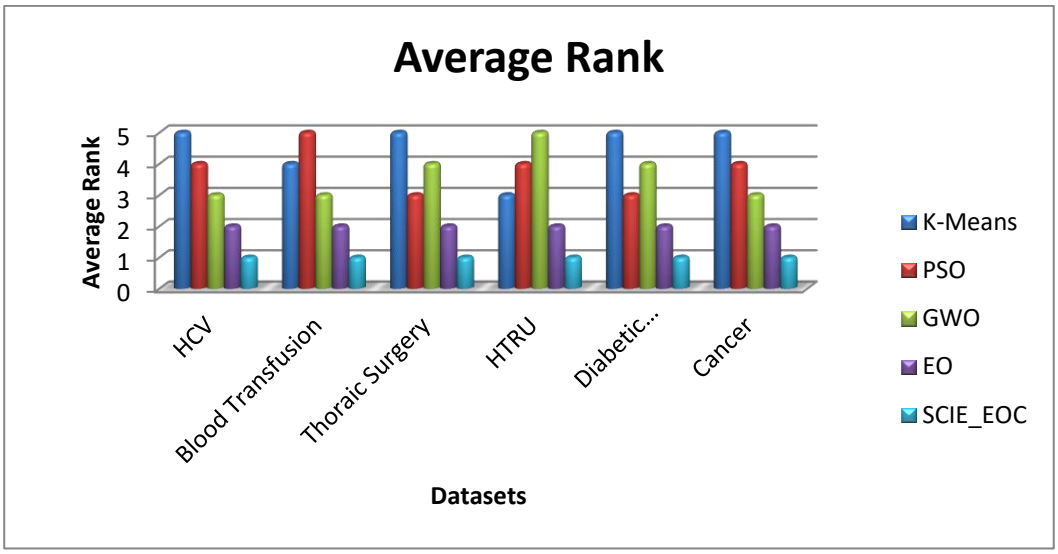

Fig. 4: Average Rank (Intra-Cluster Distance)

5.2.4. Purity Index

Purity (Pr) serves as a measure of the effectiveness of grouping (clustering) schemes, reflecting the precision of generated clustering for data components. Consequently, all components belonging to a specific group can be given to a specific group. The Purity Index (Pi) is computed using eq. (14) and eq. (15) based on purity (Pr). Maximum purity is attained with the Pi value being close to 1.

$$Pr(S_\rho) = \frac{max(|S_{\rho\mu}|)}{|S_\rho|} \qquad (14)$$

$$Pi = \sum_{\rho=1}^{R} \frac{(|S_\rho| Pr(S_\rho))}{|T|} \qquad (15)$$

Here,

$Pr(S_\rho)$ = ρ^{th} group/cluster purity.

$|S_\rho|$= ρ^{th} group/cluster Length.

$|S_{\rho\mu}|$= Overall data components of μ^{th} class consigned to ρ^{th} group/cluster.

Table 5: Purity Index

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Dataset	K-Means	PSO	GWO	EO	SCIE_EOC
HCV	0.76	0.83	0.82	0.87	0.91
Blood Transfusion	0.73	0.83	0.84	0.89	0.93
Thoraic Surgery	0.78	0.82	0.84	0.88	0.91
HTRU	0.76	0.85	0.85	0.89	0.92
Diabetic Retinopathy	0.75	0.83	0.85	0.86	0.9
Cancer	0.77	0.81	0.83	0.87	0.9

The data in Table 5 and Fig. 5 reveal that the SCIE_EOC yielded the extreme value of Pi for the six comprehensive dataset. According to Purity Index (Pi), the SCIE_EOC exceeds better than EO, GWO, PSO and K-Means by 5%, 12%, 14% and 21% respectively, across all six datasets.

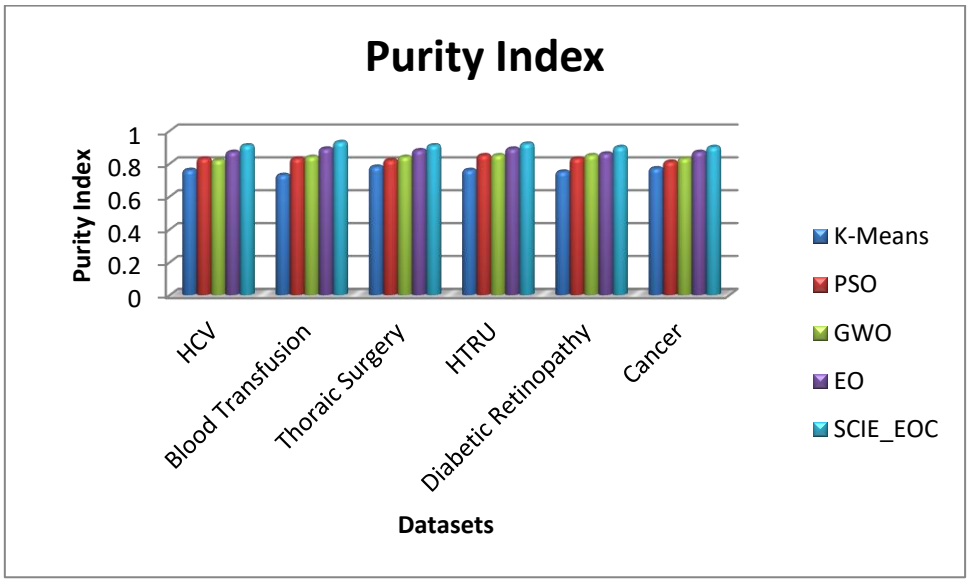

Fig. 5: Purity Index

5.2.5. F-Measure

To begin with, Precision (N) and Recall (RC) are determined to retrieve the data (eq. (16) an eq. (17)). Afterward, N and RC are integrated to assess F-Measure (F) (eq. (18) and eq. (19)).

$$N(\mu,\rho) = \frac{|S_{\rho\mu}|}{|S_\rho|} \qquad (16)$$

$$RC(\mu,\rho) = \frac{|S_{\rho\mu}|}{|S_\mu|} \qquad (17)$$

$$F(\mu,\rho) = \frac{2 \times N(\mu,\rho) \times RC(\mu,\rho)}{N(\mu,\rho) + RC(\mu,\rho)} \qquad (18)$$

$$F = \sum_{\mu=1}^{R} \frac{|S_\mu|}{|T|} max\{F(\mu,\rho)\} \qquad (19)$$

Where,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

$|S_\mu|$= μ^{th} class length.

Table 6: F-Measure

Dataset	K-Means	PSO	GWO	EO	SCIE_EOC
HCV	0.74	0.81	0.8	0.85	0.9
Blood Transfusion	0.7	0.8	0.81	0.86	0.91
Thoraic Surgery	0.75	0.79	0.82	0.85	0.89
HTRU	0.73	0.84	0.83	0.85	0.9
Diabetic Retinopathy	0.74	0.82	0.83	0.83	0.89
Cancer	0.75	0.78	0.8	0.84	0.88

According to Table 6 and Fig. 6, the SCIE_EOC determined the extreme value of F for the six comprehensive dataset. In F-Measure (F), the SCIE_EOC attains 6%, 13%, 16% and 20% greater performance than EO, GWO, PSO and K-Means respectively, across all six datasets.

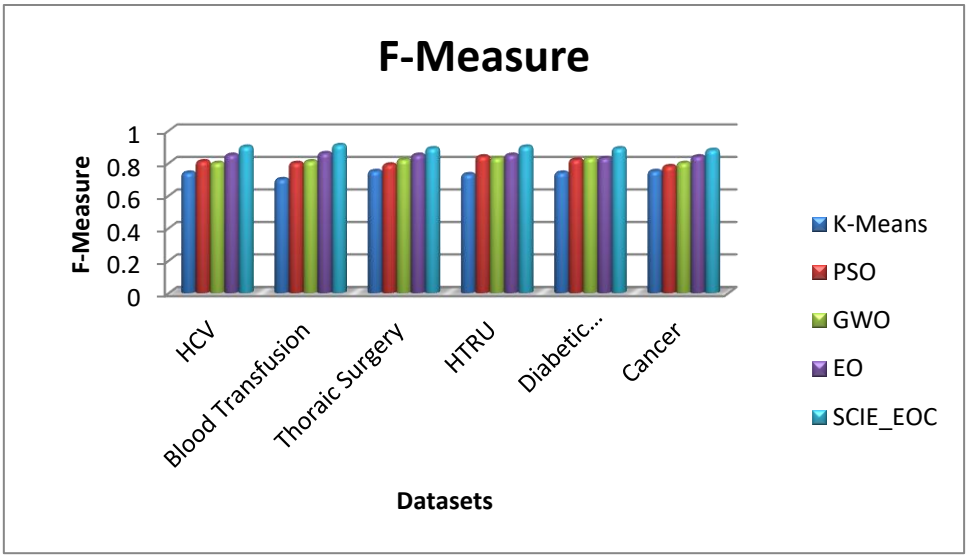

Fig. 6: F-Measure

5.2.6. Accuracy

The Accuracy (Acy) is determined based on the distribution of precise clusters (i.e., the assessment of percentage of correct decisions) and detailing the percentage of clusters in the dominant group (eq. (20)).

$$Acy = \frac{\sum_{\rho=1}^{R} S_\rho}{|T|} * 100\% \qquad (20)$$

Table 7: Accuracy

Dataset	K-Means	PSO	GWO	EO	SCIE_EOC
HCV	78	85	86	90	93
Blood Transfusion	79	83	84	87	91
Thoraic Surgery	77	81	80	86	90

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

HTRU	76	84	83	87	92
Diabetic Retinopathy	78	82	84	87	91
Cancer	79	85	86	90	93

The data in Table 7 and Fig. 7 reveal that the SCIE_EOC yielded the extreme value of Acy for the six comprehensive dataset. According to Accuracy (Acy), the SCIE_EOC exceeds better than EO, GWO, PSO and K-Means by 6%, 11%, 15% and 22% respectively, across all six datasets.

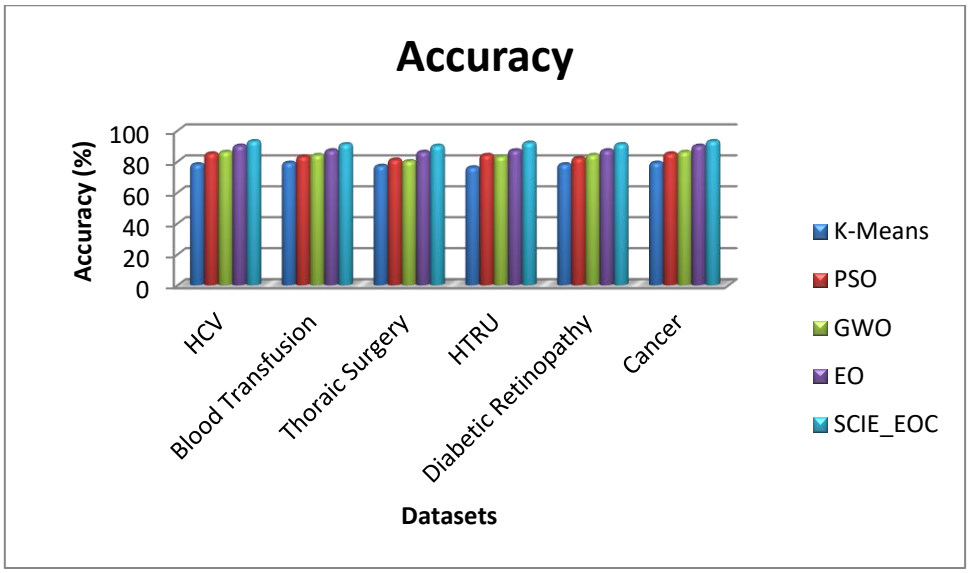

Fig. 7: Accuracy

The utilization of information entropy in conjunction with the EO scheme for clustering ensures precise and proficient dispersal of data values within the dataset. The sinusoidal chaotic is incorporated into preliminary population production of EO scheme to augment the search and utilization capabilities of elephants within clan, considerably employed for the cluster member distribution scheme. Consequently, the SCIE_EOC scheme yields optimally produced cluster centers and members of excellent quality. As a result, SCIE_EOC delivers better outcomes than EO, GWO, PSO and K-Means schemes.

5.3. SCIE_EOC Scheme Time Complexity

The time complexity is assessed by performing the clustering and enumerating the total operations. The step cost is established as 1 unit for each operation. The Full Operation Cost (FOC) is derived by applying algorithm 2 along with eq. (21) and eq. (22).

$$FOC = 1 + 1 + 1 + O^P + ITR + 1 + ITR + ITR * (T_{cn} + 1) + 2 * O^P * T_i + 3 * O^P + 1 + ITR * T_{cn} * (T_{L_n} + 1) + ITR * T_{cn} * T_{L_n} + ITR * T_{cn} * T_{L_n} + ITR * T_{cn} * T_{L_n} + ITR * (T_{cn} + 1) + ITR * T_{cn} + ITR + ITR + 1 \quad (21)$$

$$FOC = 4 * ITR * T_{cn} * T_{L_n} + 4 * ITR * T_{cn} + 2 * O^P * T_i + 6 * ITR + 4 * O^P + 6 \quad (22)$$

Gather that all parameters are alike in eq. (22) in worst case; accordingly, eq. (23) is derived as follows:

$$FOC = 4N^3 + 6N^2 + 10N + 6 \quad (23)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

In the worst case scenarios, for SCIE_EOC, EO, GWO and PSO schemes, the time complexities are $O(N^3)$ and for K-Means, it is $O(N^2)$. As a result, all schemes exhibit polynomial time complexity.

5.4. Statistical Exploration

A statistical assessment is carried out to ascertain the magnitude of noteworthy distinctions among the impacts of clustering schemes. In the context, a distribution free Friedman Assessment (FA) is conducted to explore distinctions among the set of sequential relevant variables. Every clustering scheme is equivalently productive, supporting the null hypothesis ($\emptyset_0$).

The Friedman Assessment (FA) is expressed through eq. (24).

$$FA = \frac{12 * K}{C * (C+1)} \left[\sum_{\varphi=1}^{5} (I_\varphi)^2 - \frac{C * (C+1)^2}{4} \right] \quad (24)$$

Where,

K = Total Datasets

C = Total Clustering Schemes

I_φ= φ^{th} Scheme Average Rank

The FA critical value of 2.24893 is determined from F-distribution table [35] by means of (C - 1) and (C – 1) * (K – 1) degrees of freedom lying within in the range (5 - 1) = 4 and (5 – 1) * (6 – 1) = 20 for applying 5 clustering schemes (C = 5) to 6 datasets (K = 6) having $\gamma = 0.10$ (confidence point). The estimated FA value exceeds the critical value for null hypothesis ($\emptyset_0$) rejection; otherwise $\emptyset_0$ is accepted. The FA estimated value (eq. (24)) is 20.13072 for applying 5 clustering schemes (C = 5) to 6 datasets (K = 6) having $\gamma = 0.10$. Consequently, the estimated FA value is above the crtical FA value, resulting in the rejection of $\emptyset_0$.

Thus, a post hoc evaluation is carried out using Holm scheme; in which proposed SCIE_EOC is statistically compared alongside other clustering schemes in this assessment. At the beginning, the Z-score is derived through eq. (25) and eq. (26) and subsequently, Probability (H) is determined utilizing Z-score and normal distribution table [36]. At the end, $H_\omega$value (Table 8) is derived through $\gamma/(C-\omega)$.

$$Z = \frac{I_\omega - I_\varphi}{\vartheta} \quad (25)$$

Where, 0.91

$$\vartheta = \sqrt{\frac{C * (C+1)}{6 \times K}} \quad (26)$$

Table 8: Holm Strategy Results

ω	**Approaches**	**Z-value**	**H-value**	$\gamma/(C-\omega)$	**Hypothesis**
1	K-Means	-3.8461	0.00006	0.025	Rejected
2	PSO	-3.1099	0.00071	0.033	Rejected
3	GWO	-2.9341	0.00169	0.05	Rejected

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

4	EO	-1.0989	0.13786	0.10	Accepted

According to Table 8, the $\gamma/(C-\omega)$ value exceeds the H_ω value, suggesting the rejection of the hypothsis for overall cases. Consequently, as per the above examination, the proposed SCIE_EOC and EO schemes show better clustering capabilities in comparison to GWO, PSO and K-Means schemes.

6. CONCLUSIONS AND FUTURE WORK

In this paper, the extensive use of data clustering is discussed in various sectors, highlighting its importance in decentralized scenarios for optimizing information exchange. Here, a Sinusoidal Chaotic and Information Entropy based Elephant-Herding Optimization based Clustering (SCIE_EOC) is introduced, aiming at efficiently classifying diverse datasets and enhancing data search capabilities. The SCIE_EOC focuses on achieving precise data value distribution through the integration of information entropy and sinusoidal chaotic techniques for population formation. The SCIE_EOC scheme is implemented on MATLAB 2022a software on six extensive datasets and demonstrated the remarkable efficacy, outperforming traditional schemes like K-Means, PSO, GWO and EO across multiple evaluation factors, including Root Mean Square Error, F-Measure, Standard Deviation, Purity Index, Accuracy, Intra-Cluster Distance, Time Complexity and Statistical Exploration. Overall, SCIE_EOC is shown great promise in accurately clustering diverse datasets and improving optimization performance. Going forward, the proposed scheme is predicted to forecast and validate with vast databases within the domain of bigdata. Moreover, interconnected communication is expected to provide and verify in the structure of the Internet of Things (IoT) in future.

REFERENCES

1. R. J and K. K., “Diabetes data classification using whale optimization algorithm and back propagation neural network”, Int. Res. J. Pharm., Vol. 8(11), pp. 219-222, 2017. http://dx.doi.org/10.7897/2230-8407.0811242
2. S. Bhaskaran, R. Marappan and B. Santhi, “Design and Analysis of a Cluster-Based Intelligent Hybrid Recommendation System for E-Learning Applications”, Mathematics, MDPI, Vol. 9 (107), pp. 1-21, 2021. (https://doi.org/10.3390/math9020197)
3. S. Kiruthika, EVRM. Kalaimani, R. Sudha and K. Suganthi, “ACO Feature selection and Novel Black Widow meta-heuristic Learning rate optimized CNN for Early diagnosis of Parkinson`s disease”, Turkish Journal of Computer and Mathematics Education, Vol. 12 (7), pp. 809-817, 2021.
4. D. Giordano, M. Mellia and T. Cerquitelli, “K-MDTSC: K-Multi-Dimensional Time-Series Clustering Algorithm”, Electronics, MDPI, Vol. 10 (1166), pp. 1-21, 2021. (https://doi.org/10.3390/electronics10101166)
5. A. Khare, R. Gupta and P. K. Shukla, “A Grey Wolf Optimization Algorithm (GWOA) for Node Capture Attack to Enhance the Security of Wireless Sensor Network”, International Journal of Scientific & Technology Research, Vol. 9 (3), pp. 206-209, 2020.
6. D. Sonntag, and H. J. Profitlich, “An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation”, Artificial Intelligence In Medicine, Elsevier, Vol. 93, pp-13-28, 2019. (https://doi.org/10.1016/j.artmed.2018.08.003)
7. X. Cui, Y. Li, J. Fan, T. Wang and Y. Zheng, “A Hybrid Improved Dragonfly Algorithm for Feature Selection”, IEEE Access, Vol. 8, pp. 155619-155629, 2020.
8. F. L. Vinmalar and A. K. Kombaiya, “An Improved Dragonfly Optimization Algorithm based Feature Selection in High Dimensional Gene Expression Analysis for Lung Cancer Recognition”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 9 (8), pp. 896-908, 2020.
9. T. Singh, N. Saxena, M. Khurana, D. Singh, M. Abdalia and H. Alshazly, “Data Clustering Using Moth-Flame Optimization Algorithm”, Sensors, MDPI, Vol. 21 (4086), pp. 1-19, 2021.
10. M. Daouda, and M. Mayo, “A survey of neural network-based cancer prediction models from microarray data”, Artificial Intelligence in Medicine, Elsevier, Vol. 97, pp-204-214, 2019. (https://doi.org/10.1016/j.artmed.2019.01.006)
11. X. Cui, Y. Li, J. Fan, T. Wang and Y. Zheng, “A Hybrid Improved Dragonfly Algorithm for Feature Selection”, IEEE Access, Vol. 8, pp. 155619-155629, 2020.
12. B. A. Hassan and T. A. Rashid, “Multi-disciplinary Ensemble Algorithm for Clustering Heterogeneous Datasets”, Neural Computing and Applications, pp. 1-30, 2021.
13. H. Chantar, M. Tubishat, M. Essgaer and S. Mirjalili, “Hybrid Binary Dragonfly Algorithm with Simulated Annealing for Feature Selection”, SN Computer Science, Vol. 2(295), pp. 1-11, 2021.
14. D. Giordano, M. Mellia and T. Cerquitelli, “K-MDTSC: K-Multi-Dimensional Time-Series Clustering Algorithm”, Electronics, MDPI, Vol. 10 (1166), pp. 1-21, 2021.
15. V. M. Herrera, T. M. Khoshgoftaar, F. Villanustre and B. Furht, “Random forest implementation and optimization for Big Data analytics on LexisNexis’s high performance computing cluster platform”, Journal of Big Data, pp-1-36, 2019.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

16. Y. Denga, A. Sanderb, L. Faulstichb, and K. Deneckea, “Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders”, Artificial Intelligence in Medicine, Vol. 93, pp-29-42, 2019. (https://doi.org/10.1016/j.artmed.2018.10.001)
17. N. Shahid, T. Rappon, and W. Berta, “Applications of artificial neural networks in health care organizational decision-making: A scoping review”, PLOS ONE, pp-1-22, 2019. (https://doi.org/10.1371/journal.pone.0212356)
18. K. Tadist, F. Mrabti, N. S. Nikolov, A. Zahi and S. Najah, “SDPSO: Spark Distributed PSO-based approach for feature selection and cancer disease prognosis”, Journal of Big Data, Vol. 8 (19), pp. 1-22, 2021.
19. S. Mitta, “Cognitive Computing Architectures for Machine (Deep) Learning at Scale”, Proceedings, MDPI, Vol. 1 (186), 2019.
20. T. B. Nun and T. Hoefler, “Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis”, cs.LG, pp-1-47, 2019.
21. B. A. Tama and S. Lim, “A Comparative Performance Evaluation of Classification Algorithms for Clinical Decision Support Systems”, Mathematics, MDPI, Vol. 8 (1814), pp. 1-25, 2020. (doi:10.3390/math8101814)
22. A. N. Habowski, T. J. Habowski and M. L. Waterman, “GECO: gene expression clustering optimization app for non-linear data visualization of patterns”, BMC Bioinformatics, Vol. 22 (29), pp. 1-13, 2021.
23. B. Hea, Y. Guana, and R. Dai, “Classifying medical relations in clinical text via convolutional neural networks”, Artificial Intelligence In Medicine, Elsevier Vol. 93, pp-43-49, 2019. (https://doi.org/10.1016/j.artmed.2018.05.001)
24. A. Khare, P. Shukla and S. Silakari, “Secure and Fast Chaos based Encryption System using Digital Logic Circuit”, I. J. Computer Network and Information Security, MECS, Vol. 6, pp. 25-33, 2014.
25. A. Khare, P. K. Shukla, M. A. Rizvi and S. Stalin, “An Intelligent and Fast Chaotic Encryption using Digital Logic Circuits for Ad-Hoc and Ubiquitous Computing”, Entropy, MDPI, Vol. 18 (201), pp. 1-27, 2016.
26. R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar and B. Qureshi, “An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques”, Sensors, MDPI, Vol. 20 (6076), pp. 1-23, 2020. (doi:10.3390/s20216076)
27. K. Demertzis, K. Rantos and G. Drosatos, “A Dynamic Intelligent Policies Analysis Mechanism for Personal Data Processing in the IoT Ecosystem”, Big Data and Cognitive Computing, MDPI, Vol. 4 (9), pp. 1-16, 2020. (doi:10.3390/bdcc4020009)
28. G. Talari, E. Cummins, C. McNamara and J. O`Brien, “State of the art review of Big Data and web-based Decision Support Systems (DSS) for food safety risk assessment with respect to climate change”, Trends in Food Science & Technology, Elsevier, Vol. 126, pp. 192-204, 2022.
29. Y. Yao, H. Lei, T. Gedeon and L. Zheng, “Large-scale Training Data Search for Object Re-identification”, arXiv, cs.CS, pp. 1-13, 2023.
30. J. Lei, “Association Rule Mining Algorithm in College Students` Quality Evaluation System”, Journal of Electrical and Computer Engineering, Hindawi, Vol. 2022, pp. 1-9, 2022.
31. G. Li, J. Sun, M. N. A. Rana, Y. Song, C. Liu and Z. Y. Zhu, “Optimizing High-Dimensional Functions with an Efficient Particle Swarm Optimization Algorithm”, Mathematical Problems in Engineering, Hindawi, Vol. 2020, pp. 1-10, 2020.
32. L. Abualigah, A. H. Gandomi, M. A. Elaziz, H. A. Hamad, M. Omari, M. Alshinwan and A. M. Khasawneh, “Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering”, Electronics, MDPI, Vol. 10 (101), pp. 1-30, 2021.
33. R. Srivastava, “Application of Optimization Algorithms in Clusters”, Frontiers in Chemistry, pp. 1-17, 2021.
34. S. L. M. Mosharraf and M. A. Adnan, “Improving lookup and query execution performance in distributed Big Data systems using Cuckoo Filter”, Journal of Big Data, Vol. 9 (12), pp. 1-30, 2022.
35. F Distribution Table, 2018 Mar 18. Retrieved from http://www.socr.ucla.edu/applets.dir/f_table.html.
36. Normal Distribution Table. Retrieved from http://math.arizona.edu/~rsims/ma464/standardnormaltable.pdf.