Research Article/Araştırma Makalesi

GAZİ
JOURNAL OF ENGINEERING SCIENCES

Hybridization of Meta-heuristic Algorithms with K-Means for Clustering Analysis: Case of Medical Datasets

Safa Dörterler*[a], Hatem Dumlu [b], Durmuş Ozdemir[c], Hasan Temurtas [d]

Submitted: 16.07.2023 Revised: 21.01.2024 Accepted: 09.03.2024 doi:10.30855/gmbd.0705N01

ABSTRACT

K-Means clustering is commonly used for data clustering, but it suffers from limitations such as being prone to local optima and slow convergence, particularly when handling large medical files. The literature recommends employing metaheuristic algorithms in clustering studies to address these issues. This study aims to accurately diagnose diseases in four medical datasets (Dermatology, Diabetes, Parkinson's, and Thyroid) and increase the rate of correct diagnosis of diseases. We utilized optimization algorithms to assign weights to input parameters determining diseases in these datasets, thereby improving clustering performance. Our proposed model incorporates the Crow Search Algorithm, Tree Seed Algorithm, and Harris Hawks Optimization algorithms in a hybrid structure with K-Means. We conducted statistical evaluations using performance metrics. The study demonstrated that the hybrid Harris Hawks Optimization algorithm achieved the highest accuracy rate (97.19%) among the tested algorithms on the Dermatology dataset. The hybrid Crow Search Algorithm obtained a 96.29% accuracy rate on the Thyroid dataset, while the hybrid Tree Seed Algorithm achieved a 95.32% accuracy rate on the Dermatology dataset. This study offers significant benefits, including reduced staff workload, lower test costs, improved accuracy rates, and faster test results for detecting various diseases in medical datasets.

Keywords: K-Means clustering, Metaheuristic algorithms, Disease diagnosis, Optimization, Decision support systems

[a,*] Kutahya Dumlupinar University, Engineering Faculty, Computer Engineering 43100 - Kutahya, Türkiye
Orcid: 0000-0001-8778-081X
e mail:safa.dorterler@dpu.edu.tr

[b] Kutahya Dumlupinar University, Engineering Faculty, Computer Engineering 43100 - Kutahya, Türkiye
Orcid: 0000-0002-9056-4437

[c] Kutahya Dumlupinar University, Engineering Faculty, Computer Engineering 43100 - Kutahya, Türkiye
Orcid: 0000-0002-9543-4076

[d] Kutahya Dumlupinar University, Engineering Faculty, Computer Engineering 43100 - Kutahya, Türkiye
Orcid: 0000-0001-6738-3024

*Corresponding author:
safa.dorterler@dpu.edu.tr

Kümeleme Analizi için Meta-sezgisel Algoritmaların K-Means ile Hibritlenmesi: Tıbbi Veri Kümeleri Üzerine Bir İnceleme

ÖZ

K-Means kümeleme, veri kümeleme için yaygın olarak kullanılan bir yöntemdir. Ancak özellikle büyük tıbbi verilerle çalışırken yerel optimuma takılmak ve yavaş yakınsama gibi sorunlarla karşılaşılabilir. Literatürde bu tür sorunları ele almak için kümeleme çalışmalarında metasezgisel algoritmaların kullanılmasının önerildiği görülmektedir. Bu çalışma, dört farklı tıbbi veri kümesi üzerinde (Dermatoloji, Diyabet, Parkinson ve Tiroid) hastalıkların doğru teşhisini koymayı ve hastalıkların doğru teşhis oranını artırmayı amaçlamaktadır. Bu veri kümelerindeki hastalıkları belirleyen girdi parametrelerine ağırlık atamak için optimizasyon algoritmalarını kullandık ve sonuç olarak kümeleme performansını artırdık. Önerilen modelimiz, Karga Arama Algoritması, Ağaç Tohum Algoritması ve Harris Hawks Optimizasyon algoritmalarını K-Means ile hibrit bir yapıda birleştirmektedir. Performans metrikleri kullanarak istatistiksel değerlendirmeler yaptık. Sonuçlar hibrit Harris Hawks Optimizasyon algoritmasının Dermatoloji veri kümesinde test edilen algoritmalar arasında en yüksek doğruluk oranına (%97,19) ulaştığını göstermektedir. Ayrıca hibrit Karga Arama Algoritması, Tiroid veri kümesinde %96,29 doğruluk oranı elde ederken, hibrit Ağaç Tohumu Algoritması Dermatoloji veri kümesinde %95,32 doğruluk oranı elde etmiştir. Bu çalışma, tıbbi veri kümelerinde çeşitli hastalıkları tespit etmek için daha az personel iş yükü, daha düşük test maliyetleri, gelişmiş doğruluk oranları ve daha hızlı test sonuçları gibi önemli faydalar sunmaktadır.

Anahtar Kelimeler: K-Means kümeleme, Metasezgisel algoritmalar, Hastalık teşhisi, optimizasyon, Karar destek sistemleri

To cite this article: S. Dorterler, H. Dumlu, D. Ozdemir and H. Temurtas "Hybridization of Meta-heuristic Algorithms with K-Means for Clustering Analysis: Case of Medical Datasets," *Gazi Journal of Engineering Sciences*, vol. 10, no.1, pp.1-11, 2024. doi:10.30855/gmbd.0705N01

1. Introduction

The grouping of data with similar characteristics in a dataset is called clustering [1]. Clustering operations are used in statistical data analysis, data mining, vector quantization, and data compression [2,3]. The clustering process brings problems along with its advantages [4]. The multidimensionality and large size of the data can cause time costs. In addition, when using distance-based clustering, it is difficult to determine the linked cluster in multidimensional space when it is impossible to measure the distance between clusters. Results from clustering processes can be interpreted differently due to the differences in the structure of clustering methods. [5]. Two categories, supervised and unsupervised, are typically used to categorize the classification process. The main task of supervised classification is to put unclassified data in the most appropriate class. Unsupervised classification aims to create meaningful subsets from the unclassified data in a cluster [6]. The unsupervised classification approach is used to assess the clustering process. Fraley and Raftery [7] classified clustering into two categories: hierarchical and partitional. Without notice, the top-down or bottom-up division of the number of clusters in a tree structure creates hierarchical clustering. In contrast, partitional clustering splits data into groups without regard to hierarchy and with a predetermined number of clusters. The Euclidean distance is the basis of the partitional clustering. Calculating the distance between each cluster and each point, then including that point in the cluster that minimizes this distance, is how the Euclidean distance is stated. There are many different clustering methods available. K-Means, Hierarchical clustering, and Gaussian mixture models (GMMs) for clustering are the most well-known ones [8].

Clustering analysis is used in fields such as field of medicine [9–14], machine learning [15], identification of images [16], data mining [17,18], market and consumer segmentation [19–21], biology [22], statistics [23], and pattern recognition [24]. K-Means is a popular center-based, straightforward, and quick clustering algorithm [25]. The K-Means algorithm is used in marketing [26,27], chemistry [28], geographic systems [29], meteorological phenomena [30], and social sciences [31], and it has been extensively utilized in scientific and industrial fields, particularly in the field of medicine. Liu et al. [26] performed customer classification and market analysis using the K-Means algorithm with retail company data. Heil et al. [32] classified the agricultural lands in West Africa using data based on geological and climatic parameters and the K-Means algorithm. Additionally, they demonstrated how the fuzzy K-Means algorithm is superior to the standard K-Means algorithm. Similarly, Tang et al. [28] used the K-Means algorithm to classify industrial polymers. Xiaoying et al. [33] analyzed the chemical molecules in rice and classified rice according to geographical origin. Anderson [34] used the K-Means algorithm and kernel density calculation to determine the points where traffic accidents are most intense. Bacao et al. [35] studied K-Means for self-organizing maps. Kanthan and Sujantha [30] clustered raindrops with the K-Means algorithm. Chakraborty et al. [36] suggested a methodology for weather forecasting with K-Means. Kurniawan and Fatulloh [37] used the K-Means algorithm and geographic information system data to classify the social life conditions of a city in Indonesia. Zhou et al. [38] clustered crime points using real case data. Evans et al. [39] examined the risks and negative consequences of medicines used in medicine on people. They utilized the K-Means method in order to classify medicines. They concluded that high-risk characteristics included the patient's age, weight, gender, and medicine dosage. A study conducted in the field of dentistry investigated whether psychological and social changes have an effect on the case of acute pain after surgery. The K-Means algorithm was used to cluster the patients into groups based on their psychological characteristics and symptoms. In conclusion, it was observed that patients who were female, depressed, and anxious experienced more pain after surgery [40]. In a different research project in the psychiatric field [41], various personality inventories were used for cluster analysis. By analyzing the clusters, the differences between the two groups were revealed. Thus, unknown aspects of psychopathy were tried to be discovered. Likewise, Kim et al. [42] used K-Means to divide the 888 instances submitted to the Korean emergency service into two categories. It was found that 85% of suicide attempts were impulsive, and 15% were planned based on a variety of demographic and clinical factors. Studies in the field of genetics have also used the K-Means algorithm. Shai et al. [43] revealed molecular subtypes of unknown pathological kinds and classes using the K-Means algorithm. Bertucci et al. [44] used hierarchical clustering to describe five breast cancer subtypes. According to research and observations of Ushizawa et al. [45], animal embryo gene profiles were examined using cDNA microarray, and the K-Means method was utilized to cluster genes. Doctors in the medical field diagnose patients by using a variety of tests, observations, and information about the patient's past. In the health sector, decision support systems are developed with various techniques to help doctors [46–51]. The K-Means algorithm has

the drawback of being stuck to local optima despite performing fast and effective clustering [52]. Results from the K-Means technique are based on the initial clustering reference points. In other words, the search always converges to the closest local optimum from the initial point. Researchers solve the optimum local problem using hierarchical, artificial intelligence-based, partition-based, and density-based clustering techniques. The dataset is divided into a fixed number of partitions in partition-based clustering [53]. Each point shifts its center until it is closest to its cluster center. This method is most advantageous when the dataset is homogeneous [54]. In density-based clustering, clusters are formed from areas with high data density [55]. This approach clusters data points based on their density. It is advantageous when dealing with heterogeneous datasets [54]. The partition-based clustering method focuses on specific parts, while the density-based clustering method determines clusters based on their density. Examples of the techniques used are as follows: Graph Theory [56], Artificial Neural Networks [57], Statistical methods [58], and heuristic algorithms [19, 59–64]. These techniques were used to avoid being stuck with the local optimum and to increase the clustering success rate.

In this study, the meta-heuristic algorithms Crow Search Algorithm (CSA), Harris Hawks Optimization (HHO), and Tree Seed Algorithm (TSA) are used together with the K-Means algorithm to improve clustering performance. Three new algorithms, Hybrid CSA (H-CSA), Hybrid HHO (H-HHO), and Hybrid TSA (H-TSA), were developed from the hybrid use of the algorithms. By identifying the significance of parameter values used in diagnosing diabetes, dermatology, Parkinson's, and thyroid diseases, heuristic algorithms have improved clustering success. In summary, if any parameter in a disease is more important, the importance coefficient of that parameter is increased. Thus, a more accurate diagnosis is provided. Likewise, if any parameter is less critical in a disease, the importance of this parameter is lowered. Thus, we planned to prevent misdiagnoses. Thus, the K-Means algorithm being stuck to the local optimum has been resolved. In addition, the relevant coefficients have been correctly optimized to increase the clustering success.

2. Materials and Methods

In this study, the K-Means algorithm was combined with CSA, TSA, and HHO algorithms to obtain H-CSA, H-HHO, and H-TSA algorithms that demonstrate superior clustering integrity performance. The developed model utilized four distinct medical datasets, Dermatology, Diabetes, Parkinson's, and Thyroid, as input parameters. The datasets were obtained from the UCI Machine Learning Repository and additional sources [65–69]. Figure 1 shows a summary of our methodology in the study.

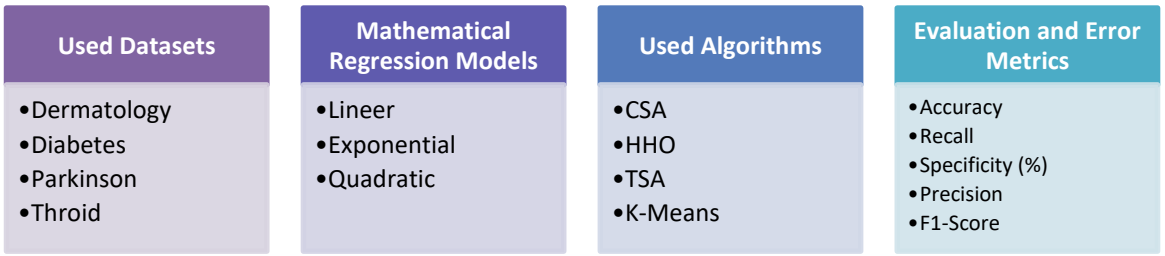

Figure 1. A brief summary of methodology

2.1. K-Means Algorithm

The K-Means algorithm, initially proposed in 1967 by James MacQueen, is a widely used and efficient clustering technique in machine learning, data exploration, and data mining [70]. The algorithm aims to group data in a dataset based on similarities and divides them into distinct clusters. The clusters are formed around the k initial cluster centers that the algorithm selects, and data are assigned to the clusters closest to these centers. Then, the centers of the clusters are recalculated, and the data is reassigned to the corresponding clusters. This step is repeated until the data distribution is corrected. This process aims to sort the data into clusters that best reflect their similarity. Equation 1 defines the objective function.

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

Where $\left\|x_i^{(j)} - c_j\right\|^2$ is the distance between $x_i^{(j)}$ and the c_j (center of cluster). The goal is to find the lowest J [8]. Thus, $x_i^{(j)}$ belongs to the c_j centered cluster for minimum J value.

A method has yet to be presented to determine the number of clusters (K) in the K-Means algorithm. This is a disadvantage besides the simplicity and popularity of the K-Means algorithm. K-Means cannot guarantee convergence to a global optimum using its iteratively optimal procedure. Additionally, the K-Means technique is susceptible to outliers and noisy data. The deformation of cluster geometries is an additional problem because it attempts to include an object in a cluster even if it is far from one [70].

2.2. Harris Hawks Algorithm

The Harris Hawk Algorithm emulates the rabbit hunting approach of the intelligent Harris hawk. Before hunting, the leader and other flock members conduct reconnaissance flights. Following prey detection, the hunting process commences. Heidari presented a mathematical model outlining these characteristics of the Harris hawk in 2019 [71].

Exploration phase: When Harris hawks roam randomly, they use two exploration strategies. These strategies are as in Equation 2. The probability value q here indicates which tactic will be in use.

$$x(t+1) = \begin{cases} x_{rand}(t) - r1\,|x_{rand}(t) - 2r_2 x(t)|, & q \geq 0.5 \\ \left(x_{rabbit}(t) - x_m(t)\right) - r_3\left(LB + r_4\,(UB - LB)\right), & q < 0.5 \end{cases} \quad (2)$$

Here x(t + 1) is the position vector of Harris in each iteration. The position vector of the prey is xrabbit(t). The current position of the hawk is x(t). r_1, r_2, r_3, r_4, and q are random numbers (0,1). The lower value and the upper value, respectively, are denoted by LB and UB. $x_{rand}(t)$ shows a hawk randomly chosen from the current population. $x_m(t)$ is the average position of the current hawk population. The average position is found using Equation [71].

$$x_m(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t) \quad (3)$$

Here, t denotes the number of iterations, and N denotes the number of hawks. After completing the exploration process, the exploitation phase is presented in Equation 4.

$$E = 2E_0\left(1 - \frac{t}{T}\right) \quad (4)$$

Here E is the total energy of the escaped prey, the prey's initial energy is E_0 is the maximum iterations number.

Exploitation phase: Four different strategies are used to simulate the exploitation phase. Soft besiege, hard besiege, soft besiege with progressive rapid dives and hard besiege with progressive rapid dives.

At the soft besiege stage, the Harris hawk makes misleading jumps so that reduce energy of prey ($r \geq 0.5, E \geq 0.5$). This soft encirclement strategy is mathematically given in Equations 5 and 6. Here, r is the chance of catching the escaped prey. E is the energy of the rabbit. $\Delta x(t)$ is the difference between the current position in the t.th iteration and the current position of the prey (rabbit). For the purpose of simulating natural rabbit movement, J is a value that changes with each iteration.

$$x(t+1) = \Delta x(t) - E|Jx_{rabbit}(t) - x(t)| \quad (5)$$

$$\Delta x(t) = x_{rabbit}(t) - x(t) \quad (6)$$

In the *Hard Besiege* strategy, the energy of the prey is considerably reduced ($r \geq 0.5, |E| \leq 0.5$). This situation is mathematically modeled as in Equation 7.

$$x(t+1) = x_{rabbit}(t) - E|\Delta x(t)| \quad (7)$$

Soft besiege with progressive rapid dives stage, it is thought that the Hawks decided their next move

according to Equation 8 before starting the soft besiege.

$$Y = x_{rabbit}(t) - E|Jx_{rabbit}(t) - x(t)| \quad (8)$$

Harris Hawks dive fast and compare to their previous dives. If the new dive situation is not suitable, the hawks continue to fast dive into their prey. A Levy Flight based motion structure is used while deciding this. Equation 9 describes this condition.

$$Z = Y + SxLF(D) \quad (9)$$

Here D is the problem size. S is a random vector of size 1xD. Y determines the position of the prey relative to its decreasing energy. Z is the variable that decides whether the hawks will attack its prey. Equation 10 gives the levy function, abbreviated LF.

$$LF(x) = 0.01x\left(\frac{\mu x \sigma}{|\mu|^{\frac{1}{\beta}}}\right), \sigma = \left[\frac{\Gamma(1+\beta)xsin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right)x\beta x2^{\left(\frac{\beta-1}{2}\right)}}\right] \quad (10)$$

Where u is the random number between v (0,1) and β is 1.5. The hawks' current locations are updated during the soft besiege phase using Equation 11.

$$x(t+1) = \begin{cases} Y \; if \; F(Y) < f(x(t)) \\ Z \; if \; F(Z) < F(x(t)) \end{cases} \quad (11)$$

Y and Z are found using Equations 8 and 9.

Hard besiege with progressive rapid dives stage, the prey lacks the energy to escape. The Harris hawk makes a hard besiege before a surprise attack to catch its prey. The hard besiege condition is found using Equation 12.

$$x'(t+1) = \begin{cases} Y' \; if \; F(Y') < f(x(t)) \\ Z' \; if \; F(Z') < F(x(t)) \end{cases} \quad (12)$$

Where Y' and Z' are found by Equations 13 and 14.

$$Y' = x_{rabbit}(t) - E|Jx_{rabbit}(t) - x_m(t)| \quad (13)$$

$$Z' = Y' + SxLF(D) \quad (14)$$

2.3. Tree Seed Algorithm

Kiran [72] introduced the Tree Seed Algorithm (TSA) in 2015, a new metaheuristic optimization algorithm that addresses continuous optimization problems. The TSA is based on the inherent relationship between trees and seeds in nature. Tree seeds are spread in the soil and grow into trees over time [73]. The position of trees and seeds indicates potential solutions for persistent issues when tree soils are considered in the research field [74]. As a result, the significance of seed sites in the formation of trees has grown. The search space is described by two equations. The first is the procedure for producing seeds for the best tree population placement. This strengthens the algorithm's ability to perform local searches. For the purpose of creating a new seed, the other equation uses two alternative tree places [72].

$$S_{i,j} = T_{i,j} + (a_{i,j})x(B_j - T_{r,j}) \quad (15)$$

$$S_{i,j} = T_{i,j} + (a_{i,j})x(T_{i,j} - T_{r,j}) \quad (16)$$

where, $S_{i,j}$ is jth dimension of ith seed that will be produced ith tree,$T_{i,j}$ is the jth dimension of ith tree, B_j is the jth dimension of best tree location obtained so far, $T_{r,j}$ is the jth dimension of rth tree randomly selected from the population, the scaling factor is α which is randomly produced in range of [−1, 1] and i and r are seperate indices.

Equation 17 is used to generate the initial tree locations, which are potential solutions to the optimization problem, at the beginning of the TSA search.

$$T_{i,j} = L_{j,min} + r_{i,j}\,(H_{j,max} - L_{j,min}) \tag{17}$$

where, $L_{j,min}$ is the lower bound of the search space, $H_{j,max}$ is the higher bound of the search space and $r_{i,j}$ is a random number produced for each dimension and location, in range of [0, 1].

Equation 18 is used to choose the best solution from the population for minimization.

$$B = min\{f(\vec{T}_i)\} i = 1,2,\dots,N \tag{18}$$

N is the total population of trees.

The maximum number of function evaluations (Max_FEs) is chosen as the termination condition, and it is set using the function's dimensionality given in Equation 19.

$$Max_{FEs} = Dx10.000 \tag{19}$$

2.4. Crow Search Algorithm

Crows live in flocks and have a powerful memory [75]. They store the food they find and return to their hiding place when needed. Crows can also follow each other to learn the location of food stores and steal each other's food. Crows can take precautions against this by flying to different places to confuse other crows. Inspired by this intelligent behavior of crows, Askarzadeh developed the Crow Search Algorithm to solve the optimization problem [75]. The algorithm consists of 4 steps.

Step 1: Initial values are given to algorithm parameters. Then, N crows are randomly placed in the d-dimensional search space. In the initialization phase, all crows store their food in the positions in which they are randomly placed. The fitness values of all positions are calculated, and their memory holding their best position is initialized.

Step 2: Each crow randomly chooses a crow and follows it to find the crow's food. New positions are created depending on whether the jth crow knows or does not know that it is being followed by the ith crow Equation 20.

$$x^{i,itr+1} = \begin{cases} x^{i,itr} + r_i \times fl^{i,itr} \times (m^{j,itr} - x^{i,itr}), & r_i \geq AP^{j,itr} \\ a, & otherwise \end{cases} \tag{20}$$

Here $x^{i,itr}$ denotes the position of the ith crow in itrth iteration. r_i is a randomly distributed random number between 0-1, m (memory) is the variable that holds the best position of the crow, and a is the random position. fl (flight length), the algorithm's adjustment parameters are called AP (awareness probabilities). fl is the range at which the crow can fly, and AP represents the probability of the crow noticing that it is being followed.

Step 3: The feasibility of the new locations found is checked. The crows' positions may be changed as necessary to create new ones. Otherwise, the crow position will not be changed.

Step 4: All positions' fitness values are calculated, and the memories of the crows are updated according to Equation 21.

$$m^{i,itr+1} = \begin{cases} x^{i,itr+1} & f(x^{i,itr+1}) is\ better\ than\ f(m^{i,itr}) \\ m^{i,itr} & otherwise \end{cases} \tag{21}$$

Here $f(x^{i,itr+1})$ shows the fitness value of the ith crow. If the fitness value of the new position is better than the fitness value of the location that was previously memorized, the update takes place according to the new position. If the stopping requirement is not met, iteration continues. The position of the crow with the best fitness value is selected, and the algorithm is terminated.

2.4. Proposed model

The K-Means algorithm works effectively when the distribution of objects in the dataset is normally distributed. However, if the distribution of objects in the dataset is non-normally distributed, the algorithm may become ineffective. Similarly, suppose there are significant or minimal differences between the properties of the objects in the data set. In that case, the K-Means algorithm will not work effectively and may not give what is expected [76,77]. K-Means approach and heuristic algorithms (CSA, HHO, and TSA) were combined in this study to improve clustering performance. Meta-heuristic algorithms were utilized for weighting various health data used in diagnosing Dermatology, Diabetes, Parkinson's, and Thyroid ailments. In summary, high or low-impact data in determining the disease were found with meta-heuristic algorithms, increasing diagnostic success and making more accurate diagnoses. Thus, the classical K-Means algorithm was prevented from being stuck at local optimum points. Clustering success was increased by optimizing the disease coefficients according to the degree of importance, thus preventing misdiagnoses. First, a W pool is created to rate all dimensions of health data.

$$W = [w_1, w_2, \dots, w_m] \tag{22}$$

W are the weight values that the optimization algorithm will optimize. Each w value is weighted by the optimization algorithm according to its importance in the diagnosis of the disease. All dimensions (X_n) are multiplied by a weight parameter (W_m). Thus, the coefficient of the importance of the relevant data (dimension) in determining the disease is found.

$$X_{n,m} = W_m * X_{n,m}, \quad n = 1, \dots, N, \quad m = 1, \dots, M, \quad n \leq N, \quad m \leq M, \quad M < N \tag{23}$$

where N represents the total amount of data and M represents the total number of data attributes. To generate a new $X_{n,m}$ pool, each characteristic is multiplied by W_m. After that, Equation 24 is used to determine the centers of all the data in this pool.

$$C_k = \frac{\sum_{n=1}^{N} a_{n,k} * X_n}{\sum_{n=1}^{N} a_{n,k}}, a_{n,k} = \begin{cases} 1 & y_n = k \\ 0 & y_n \neq k' \end{cases} \qquad k = 1, 2, \dots, K \tag{24}$$

Here K represents the number of clusters, y_n represents the data set, and a represents a variable consisting of 1 or 0. The actual value of the data is compared with the first set *(k.)*. If a match is achieved as a result of the comparison, a takes the value of 1, and the *(k.)* elements of the relevant set are added. This way, the cluster's center point in the k row is found. The center points of all clusters are found in this way. Then, using Equation 25, the class of the data is found according to the center points.

$$f_n = k, \ if \ \min(|x_n - c_1|, |x_n - c_2|, \dots, |x_n - c_K|,) \tag{25}$$

f_n is found values that represent data sets. The data is assigned to the f_n variable to belong to whichever center point is closer. Then, the weight values that give the optimum result are found with the proposed meta-heuristics by processing the available data. The objective function used is shown in Equation 26.

$$\max f(v) = {}^{100}/_{N} * \sum_{n=1}^{N} b_n, \qquad b_n = \begin{cases} 1 & y_n = f_n \\ 0 & y_n \neq f_n \end{cases} \tag{26}$$

Here, N is the number of data, and b is the variable that takes the value 0 or 1. If the found value and the actual value are equal, b is given the value 1; otherwise, it is given the value 0. These steps continue until the number of iterations is completed. Then, the weight values that give the optimum fitness value are found. An illustration of these procedures is shown in a flow chart in Figure 2.

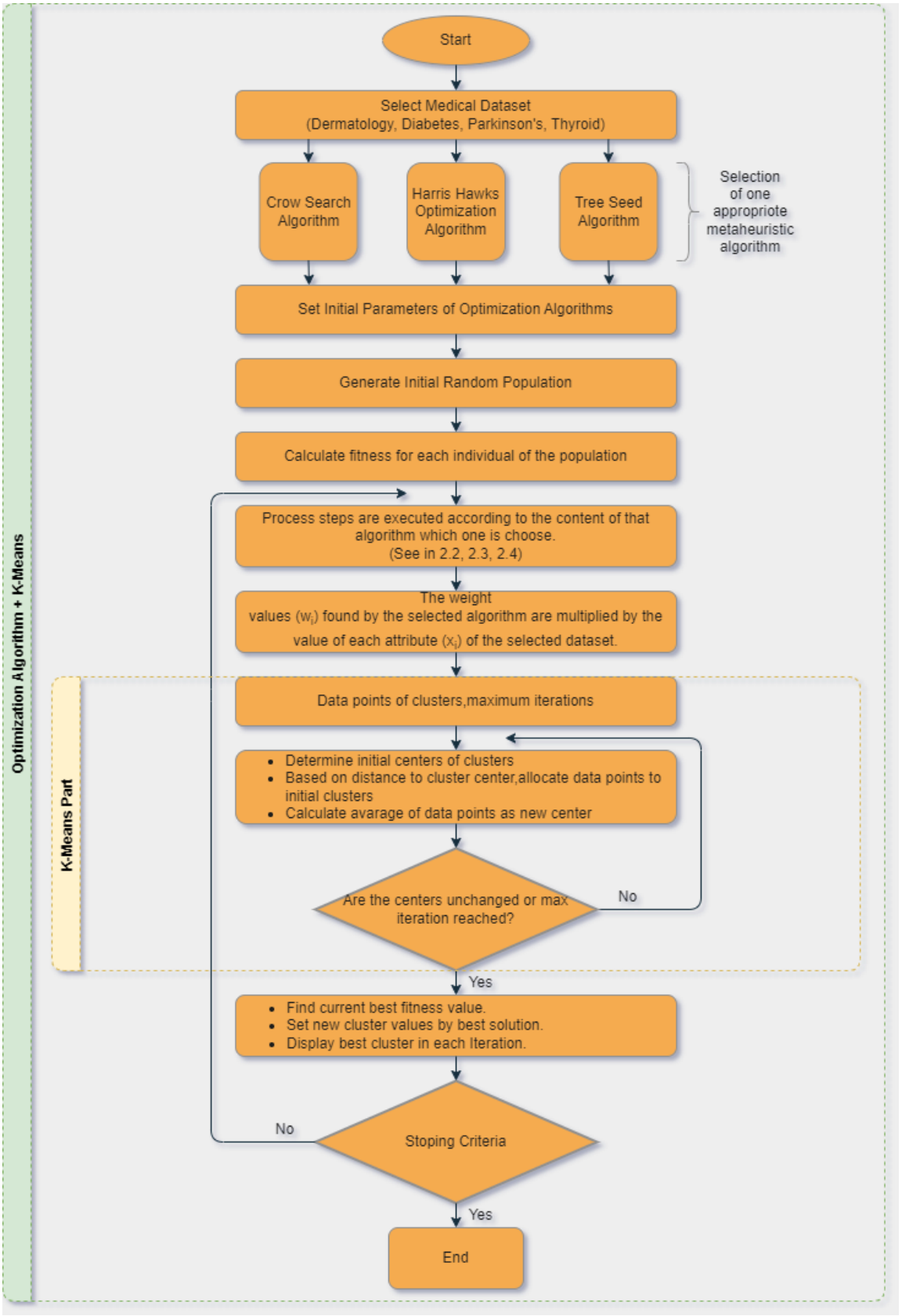

Figure 2. Flow chart of conceptual approach

2.6. Dataset

Four distinct medical datasets were employed in this research: Dermatology, Diabetes, Parkinson's, and Thyroid. The datasets were obtained from several sources, including the UCI Machine Learning Repository, and 70% of the data was allocated for training purposes, while the remaining 30% was reserved for testing [65–69].

2.6.1 Dermatology dataset

The dermatology dataset was obtained from UCI Machine Learning Repository [65,69]. There are 34 attributes in the dataset. There are 358 sample data and six different classes (Pityriasis Rubra Pilaris, Psoriasis, Seborrheic Dermatitis, Pityriasis Rosea, Lichen Planus, Chronic Dermatitis).

2.6.2. Diabetes dataset

The diabetes dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases [66]. The dataset has eight attributes. These are age, blood pressure, pregnancy, skin thickness, glucose, insulin, BMI and diabetes pedigree function. This dataset contains 768 records. It is divided into two results: diabetic or not.

2.6.3. Parkinson dataset

This dataset is taken from the UCI Machine Learning Repository [65,67]. It consists of different biomedical sound measures obtained from 195 Parkinson's disease patients. This dataset contains 22 attributes, and there are two groups of outcomes: "1" if Parkinson's disease is present and "0" if not.

2.6.4. Thyroid function dataset

The thyroid dataset was taken from the UCI Machine Learning Repository [65,68]. This dataset has 7200 records and 20 attributes, including sex, age, query_on_thyroxine, on_thyroxine, on_antithyroid_medication, pregnant, sick, thyroid_surgery, I131_treatment, query_hyperthyroid, query_hypothyroid, tumor, lithium, psych, goiter, hypopituitary, T3, TSH, T4U, TT4. There are three classes in the result set. These are not-hypothyroid (normal), subnormal function and hyperfunction.

Table 1. Datasets of study

Type of Feature	Datasets	Inputs	Classes	Instances
Numeric	Dermatology Dataset	34	6	358
Numeric	Diabetes Dataset	8	2	768
Numeric	Parkinson Dataset	22	2	195
Numeric	Thyroid Function Dataset	20	3	7200

2.7. Statistical performance metrics

The models used in the current study were evaluated with performance measurement metrics. Formulas of performance metrics are given in Equations 27, 28, 29, 30, and 31 [78–80].

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \tag{27}$$

$$Recall = \frac{TP}{TP+FN} \tag{28}$$

$$Specificity = \frac{TN}{TN+FP} \tag{29}$$

$$Precision = \frac{TP}{TP+FP} \tag{30}$$

$$F1\ Score = 2 \times \frac{Presicion \times Recall}{Precision + Recall} \quad (31)$$

Some abbreviations are used in the equations. True Positive (TP) represents a sick person as sick, False Positive (FP) represents a healthy as sick (which is incorrect), True Negative (TN) represents a healthy person as not sick, False Negative (FN) represents healthy to the person who is sick.

3. Experimantal Results

In this study, CSA, HHO, TSA, and K-Means algorithms were hybridized to create H-CSA, H-HHO, and H-TSA algorithms and were used to diagnose four different diseases. The algorithm was run on a computer with an Intel Core i7, 2.4 GHz CPU, and 8 GB RAM with Windows® 11 operating system, using MATLAB 2022b programming language. The parameter settings used in the algorithms are given in Table 2.

Table 2. Parameter settings

Algorithm	Parameter Settings	Iteration Number	Population Number
CSA	*AP=0.1; *FL=2	100	40
HHO	*ß=1.5	100	40
TSA	*ST=0.1	100	40

* AP (Awareness Probability), FL (Flight Length), ß (Beta), ST (Search Tendency)

The termination criterion of the proposed hybrid model is the number of iterations. The proposed model runs for 100 iterations and then terminates and finds the weight values that give the best result. In this way, the algorithms are run 20 times, and the weight values that give the best fitness value are found. The weight values obtained for the dermatology, diabetes, Parkinson's, and thyroid datasets are presented in Table 3. Furthermore, Table 4 presents the statistical results of mean runtime, Standard Deviation, Average Fitness, Worst Fitness, and Best Fitness values. Each dataset has been evaluated separately, and better results produced are shown in bold.

The hybrid models proposed for each dataset are separately evaluated in Table 4, and the values that yield superior results are shown in bold. The H-HHO algorithm found more successful fitness values in the Dermatology dataset than other algorithms regarding the worst and best fitness values. On the other hand, the H-CSA algorithm produced better values than other algorithms in terms of average fitness, worst fitness, and mean runtime values. In terms of standard deviation, H-TSA produced better results. H-CSA and H-HHO Diabetes found the best fitness values in the dataset. The best values in average fitness and standard deviation were obtained from the H-TSA algorithm. In the Parkinson's dataset, the H-TSA algorithm found the best fitness with 81.03%. Similarly, the most successful algorithm regarding the average fitness value is H-TSA. For the Thyroid dataset, the H-HHO algorithm gave the best values in Standard Deviation, Average Fitness, Best Fitness, and Worst Fitness. The H-CSA algorithm was found to have the lowest average runtime when evaluating the statistical results in terms of mean time. Table 5 shows the fitness values and runtime results of the K-Means Algorithm for each data set separately.

Analyzing the results in Tables 4 and 5, it is evident that the H-CSA, H-HHO, and H-TSA models outperform K-Means in terms of best fitness values. However, hybrid algorithms use two algorithms, so the proposed models run slower than the K-Means algorithm. However, the high achievements show that hybrid algorithms can make this speed difference tolerable and advantageous, especially when problem complexity and dataset characteristics are considered. In addition, the complexity matrices of the algorithms using the test dataset are presented and statistical performance results are given. Table 6 shows the values of the results for the Dermatology test dataset.

Table 3. Coefficient of Dermatology, Diabetes, Pakinson's, and Thyroid dataset

	Dermatology			**Diabetes**			**Parkinson's**			**Thyroid**		
	CSA	**HHO**	**TSA**	**CSA**	**HHO**	**TSA**	**CSA**	**HHO**	**TSA**	**CSA**	**HHO**	**TSA**
W1	0.99955	0.68342	0.02798	0.29578	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
W2	0.99926	0.48179	0.48179	0.74054	0.28513	0.67698	0.0	0.03168	0.0	0.0	0.0	0.06567
W3	1.0	1.0	0.37077	0.37931	0.0	0.34710	0.0	0.0	0.0	0.0	0.0	0.03777
W4	0.99963	1.0	0.62532	0.52302	0.0	0.32596	0.0	0.0	0.27688	0.0	0.0	1.0
W5	0.95734	1.0	0.66528	0.00028	0.0	0.05182	0.0	1.0	1.0	0.0	0.0	0.27827
W6	0.99987	1.0	0.84918	0.99960	0.73997	0.77061	1.0	0.0	1.0	0.0	0.0	0.07903
W7	0.88161	1.0	1.0	0.60541	0.28964	0.72123	0.0	1.0	0.30304	0.0	0.0	0.34065
W8	0.95635	1.0	0.86321	0.98806	0.39522	0.83784	1.0	1.0	0.0	0.0	0.0	0.94969
W9	0.99971	1.0	0.83405	N/A	N/A	N/A	0.0	0.0	0.90080	0.00136	0.0	1.0
W10	0.99922	0.53874	0.15785	N/A	N/A	N/A	0.0	0.0	0.09686	0.47101	0.0	0.76311
W11	0.99999	1.0	0.69349	N/A	N/A	N/A	1.0	0.18872	0.21580	0.0	0.0	0.17567
W12	0.99974	0.30981	1.0	N/A	N/A	N/A	0.0	1.0	0.21538	0.0	0.0	0.21051
W13	0.99955	0.65322	0.57133	N/A	N/A	N/A	0.0	0.0	0.11129	0.0	0.0	0.0
W14	0.99967	1.0	0.83357	N/A	N/A	N/A	0.80071	1.0	0.63166	0.0	0.0	0.84297
W15	1.0	1.0	0.99547	N/A	N/A	N/A	1.0	1.0	0.41264	1.0	0.0	0.42849
W16	0.99962	0.61753	0.46498	N/A	N/A	N/A	0.0	0.0	0.01859	0.04365	0.0	0.0
W17	0.99988	0.70193	0.52315	N/A	N/A	N/A	0.0	1.0	0.20289	0.0	0.00513	0.79203
W18	0.99925	0.29884	0.95313	N/A	N/A	N/A	1.0	0.0	0.47236	0.0	0.0	0.95175
W19	0.99922	0.85027	0.41822	N/A	N/A	N/A	1.0	0.48010	0.07504	0.0	0.0	0.97276
W20	1.0	1.0	0.11431	N/A	N/A	N/A	1.0	1.0	0.75772	0.18877	0.0	0.47394
W21	0.99916	1.0	0.12406	N/A	N/A	N/A	1.0	1.0	0.16217	1.0	0.0	0.99125
W22	0.02495	1.0	0.85538	N/A	N/A	N/A	1.0	0.21911	0.18336	N/A	N/A	N/A
W23	1.0	1.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W24	0.02037	1.0	0.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W25	0.99949	1.0	0.19679	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W26	0.74203	1.0	1.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W27	0.99938	1.0	0.58594	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W28	0.99986	1.0	0.80334	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W29	0.55448	0.19703	1.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W30	0.99974	1.0	0.82548	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W31	0.99993	1.0	0.76806	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W32	0.99995	0.43120	0.36214	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W33	0.0	1.0	0.42506	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
W34	0.0	0.04693	0.05681	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 4. Comparison of fitness values and running times of hybrid algorithms

Dermatology Dataset	**Diabetes Dataset**	**Parkinson's Dataset**	**Thyroid Dataset**

	H-CSA	H-HHO	H-TSA	H-CSA	H-HHO	H-TSA	H-CSA	H-HHO	H-TSA	H-CSA	H-HHO	H-TSA
Best Fitness	96.2616	**97.1962**	95.3271	**73.4782**	**73.4782**	73.0434	79.3103	79.3103	**81.0344**	96.2962	**96.9907**	90.3703
Worst Fitness	**93.4579**	**93.4579**	**93.4579**	**69.5652**	**69.5652**	**69.5652**	67.2413	**70.6896**	**70.6896**	86.2962	**94.9537**	87.3611
Average Fitness	**95.9532**	95.0934	94.4859	71.4565	71.5217	**71.6086**	72.5000	74.3965	**77.3275**	90.3842	**95.5671**	88.0092
Standard Deviation	1.0677	1.0879	**0.7365**	1.2238	1.2418	**0.8707**	3.6883	**2.3913**	3.0266	3.9746	**0.5241**	1.0684
Mean Runtime(ms)	**1.6550**	3.0765	11.8207	**1.0888**	1.9590	7.3823	**0.4759**	0.8472	3.2175	**12.4584**	22.1137	87.0473

Table 5. Fitness values and runtime scores of the K-Means Algorithm for medical data sets

	Dermatology Dataset	Diabetes Dataset	Parkinson's Dataset	Thyroid Dataset
Best Fitness	0.051248	0.040272	0.037009	0.022738
Mean Runtime (ms)	44.8598%	61.7391%	68.9655%	43.4259%

Table 6. Performance results of hybrid algorithms and K-Means algorithm for Dermatology dataset (Acc:Accuracy, Rcl:Recall, Spc:Specificity, Pre:Precision, F1s:F1-Score)

H-CSA
Train Score: 99.6015%
Test Score: 96.2616%

	Psoriasis	Seboreic Dermatitis	Lichen Planus	Pityriasis Rosea	Chronic Dermatitis	Pityriasis Rubra Pilaris
Acc(%)	1.0	0.9626	1.0	0.9626	1.0	1.0
Rcl (%)	1.0	0.8500	1.0	0.9090	1.0	1.0
Spc (%)	1.0	0.9885	1.0	0.9687	1.0	1.0
Pre(%)	1.0	0.9444	1.0	0.7692	1.0	1.0
F1s(%)	1.0	0.8947	1.0	0.8333	1.0	1.0

H-HHO
Train Score: 98.8047%
Test Score: **97.1962**%

	Psoriasis	Seboreic Dermatitis	Lichen Planus	Pityriasis Rosea	Chronic Dermatitis	Pityriasis Rubra Pilaris
Acc(%)	1.0	0.9813	0.9906	0.9719	1.0	1.0
Rcl (%)	1.0	0.9500	0.9545	0.9090	1.0	1.0
Spc (%)	1.0	0.9885	1.0	0.9791	1.0	1.0
Pre(%)	1.0	0.9500	1.0	0.8333	1.0	1.0
F1s(%)	1.0	0.9500	0.9767	0.8695	1.0	1.0

H-TSA
Train Score: 100.0%
Test Score: 95.3271%

	Psoriasis	Seboreic Dermatitis	Lichen Planus	Pityriasis Rosea	Chronic Dermatitis	Pityriasis Rubra Pilaris
Acc(%)	1.0	0.9532	1.0	0.9532	1.0	1.0
Rcl (%)	1.0	0.8000	1.0	0.9090	1.0	1.0
Spc (%)	1.0	0.9885	1.0	0.9583	1.0	1.0
Pre(%)	1.0	0.9411	1.0	0.7142	1.0	1.0
F1s(%)	1.0	0.8648	1.0	0.8000	1.0	1.0

K-Means Algorithm
Train Score: 45.0199%
Test Score: 44.8598%

	Psoriasis	Seboreic Dermatitis	Lichen Planus	Pityriasis Rosea	Chronic Dermatitis	Pityriasis Rubra Pilaris
Acc(%)	0.8411	0.7102	0.6635	0.9158	0.9158	0.8504
Rcl (%)	0.4687	0.2500	0.5454	0.2727	0.3076	1.0
Spc (%)	1.0	0.8160	0.6941	0.9895	1.0	0.8367
Pre(%)	1.0	0.2380	0.3157	0.7500	1.0	0.3600
F1s(%)	0.6382	0.2439	0.400	0.4000	0.4705	0.5294

The test dataset results for the dermatology classification of disease types were evaluated separately for each disease type. Algorithms that make more successful classification are shown in green for Psoriasis, blue for Seborrheic Dermatitis, red for Lichen Planus, purple for Pityriasis Rosea, orange for Chronic Dermatitis, and brown for Pityriasis Rubra Pilaris. The H-HHO algorithm was more successful than the H-CSA and H-TSA algorithms, with a 97% success rate for the Dermatology dataset. The H-HHO algorithm has the highest accuracy rate in 5 disease types: Seborrheic Dermatitis, Psoriasis, Chronic Dermatitis, Pityriasis Rosea, and Pityriasis Rubra Pilaris. In Lichen Planus disease, it has a very high accuracy rate, with a value of 0.9719, close to 1. H-CSA and H-TSA algorithms have an accuracy score of 1 in diagnosing Lichen Planus disease. It has succeeded in diagnosing Psoriasis, pityriasis rubra, and chronic dermatitis diseases with an accuracy of 1.0 in all three algorithms. When the test metrics are evaluated, although the H-HHO algorithm finds relatively more successful scores in 6 disease types than the H-CSA and H-TSA algorithms, the test metrics of all three optimization algorithms are high and can be used successfully to diagnose this disease. The test score of the K-Means algorithm was examined, and a low success rate of approximately 45% was found. The success rate increased significantly when the K-Means algorithm was used in a hybrid way with optimization algorithms. It was approximately 95% for all three optimization algorithms. In Figure 3, matrices of dermatology dataset confusion of the four algorithms are given.

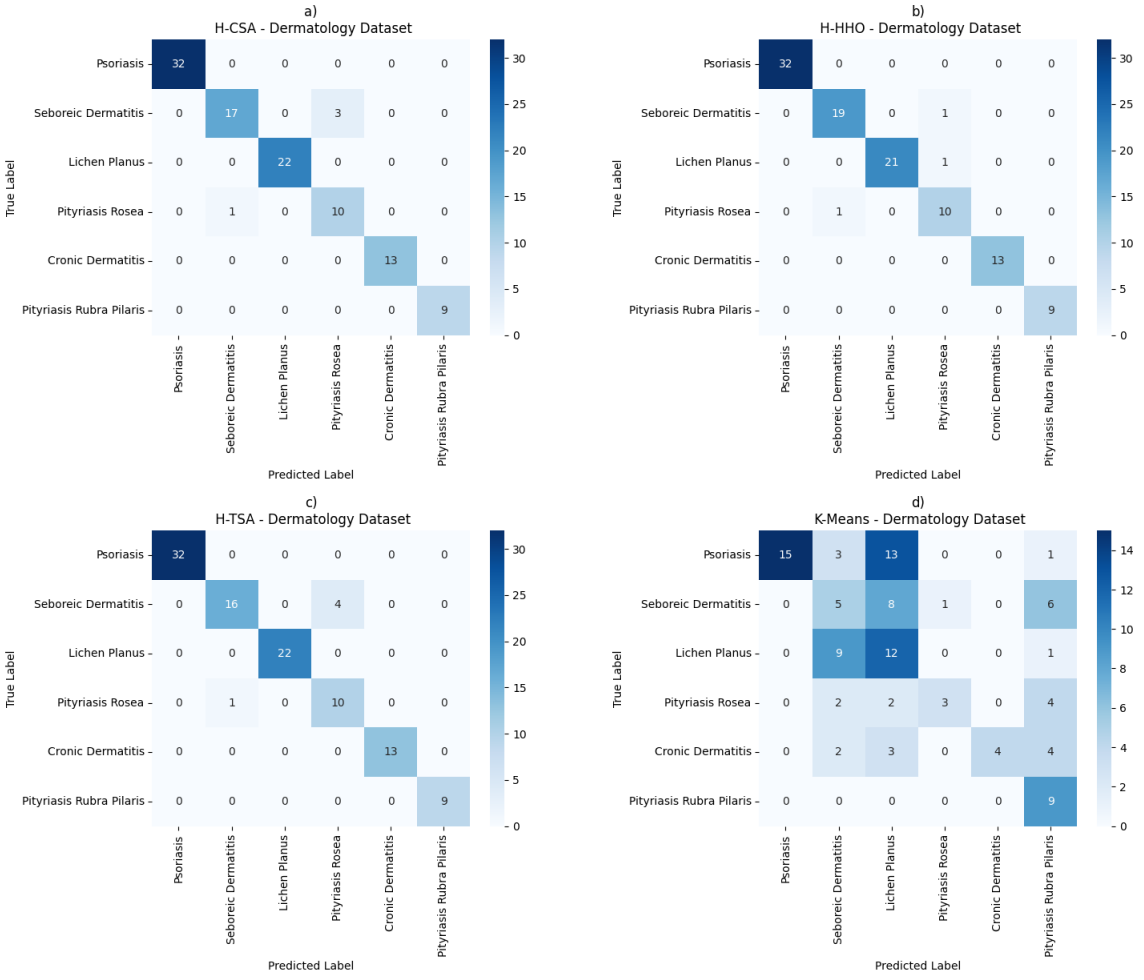

Figure 3. Confusion matrix results of hybrid algorithms and K-Means algorithm for Dermatology dataset

Compared to the optimization algorithms, the disease classification performance of the K-Means algorithm is weak. Hybrid algorithms achieved high performance in classifying all disease types compared to K-Means. Diabetes test results are given in Table 7 as follows. Algorithms with more successful classifications are shown in bold.

Table 7. Performance results of hybrid algorithms and K-Means algorithm for Diabetes dataset

	H-CSA	H-HHO	H-TSA	K-Means Algorithm
Algorithm	Train Score: 81.7518% Test Score: 79.3103%	Train Score: 81.7518% Test Score: 79.3103%	Train Score: 81.7518% Test Score: **81.0344**%	Train Score: 73.7226% Test Score: 68.9655%
Accuracy (%)	0.7931	0.7931	**0.8103**	0.6896
Recall (%)	**1.0**	0.8333	0.8333	0.6875
Specificity (%)	0.7391	0.7826	**0.8043**	0.6904
Precision (%)	0.5	0.5	**0.5263**	0.4583
F1-Score (%)	**0.6666**	0.6250	0.6451	0.5500

H-CSA and H-HHO algorithms achieved the same result in classifying diabetes disease and produced more successful results than H-TSA. The proposed H-CSA and H-HHO models correctly classified diabetes with approximately 74% accuracy. Compared to the other two optimization algorithms, the H-TSA algorithm is also classified with very close accuracy. Accordingly, the H-TSA model is also classified with approximately 73% accuracy. However, the H-CSA and H-HHO algorithms achieved higher performance than the H-TSA algorithm according to accuracy, specificity, precision, and F1-score values. The hybrid algorithms performed about 20% more accomplished classifications than the K-Means algorithm. In Figure 4, confusion matrices taken using the diabetes dataset are shown.

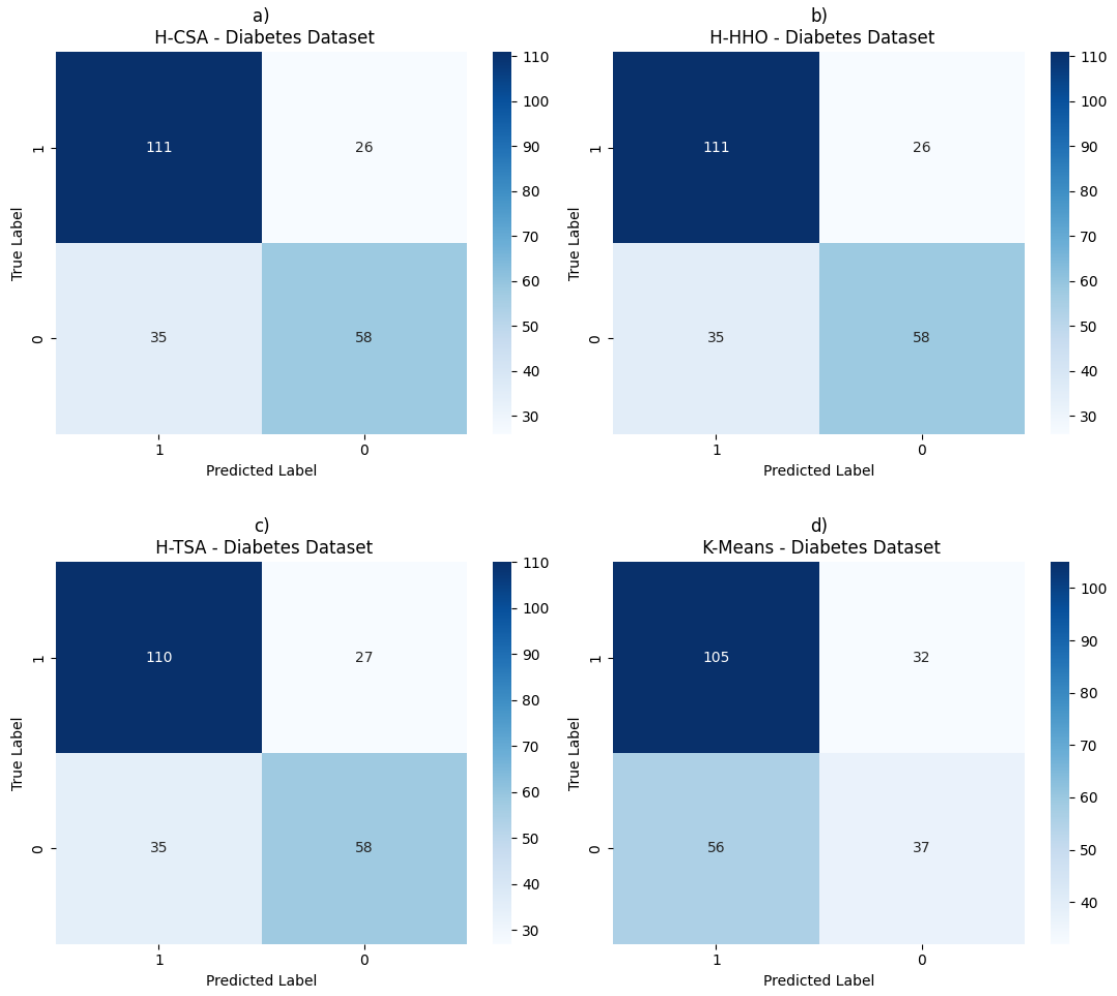

Figure 4. Confusion matrix results of hybrid algorithms and K-Means algorithm for Diabetes dataset

Parkinson's Dataset test results are given in Table 8. Algorithms with more successful results are shown in bold.

Table 8. Performance results of hybrid algorithms and K-Means algorithm for Parkinson's dataset

	H-CSA	H-HHO	H-TSA	K-Means Algorithm
Algorithm	Train Score: 81.7518% Test Score: 79.3103%	Train Score: 81.7518% Test Score: 79.3103%	Train Score: 81.7518% Test Score: **81.0344**%	Train Score: 73.7226% Test Score: 68.9655%
Accuracy (%)	0.7931	0.7931	**0.8103**	0.6896
Recall (%)	**1.0**	0.8333	0.8333	0.6875
Specificity (%)	0.7391	0.7826	**0.8043**	0.6904
Precision (%)	0.5	0.5	**0.5263**	0.4583
F1-Score (%)	**0.6666**	0.6250	0.6451	0.5500

The H-TSA algorithm achieved the highest performance in the Parkinson's dataset, with an accuracy of about 81%. In addition, the H-TSA algorithm received the highest scores in the specificity and precision metrics. The H-TSA algorithm was followed by the H-CSA and H-HHO algorithms with a 79% accuracy score. According to the recall and F1-score metrics, the H-CSA algorithm is ahead of other algorithms. In summary, all three hybrid algorithms achieved a successful classification with scores close to each other. On the other hand, the K-Means algorithm lags far behind the hybrid algorithms, with an accuracy rate of about 68%. In Figure 5, confusion matrices obtained using the Parkinson's dataset are given.

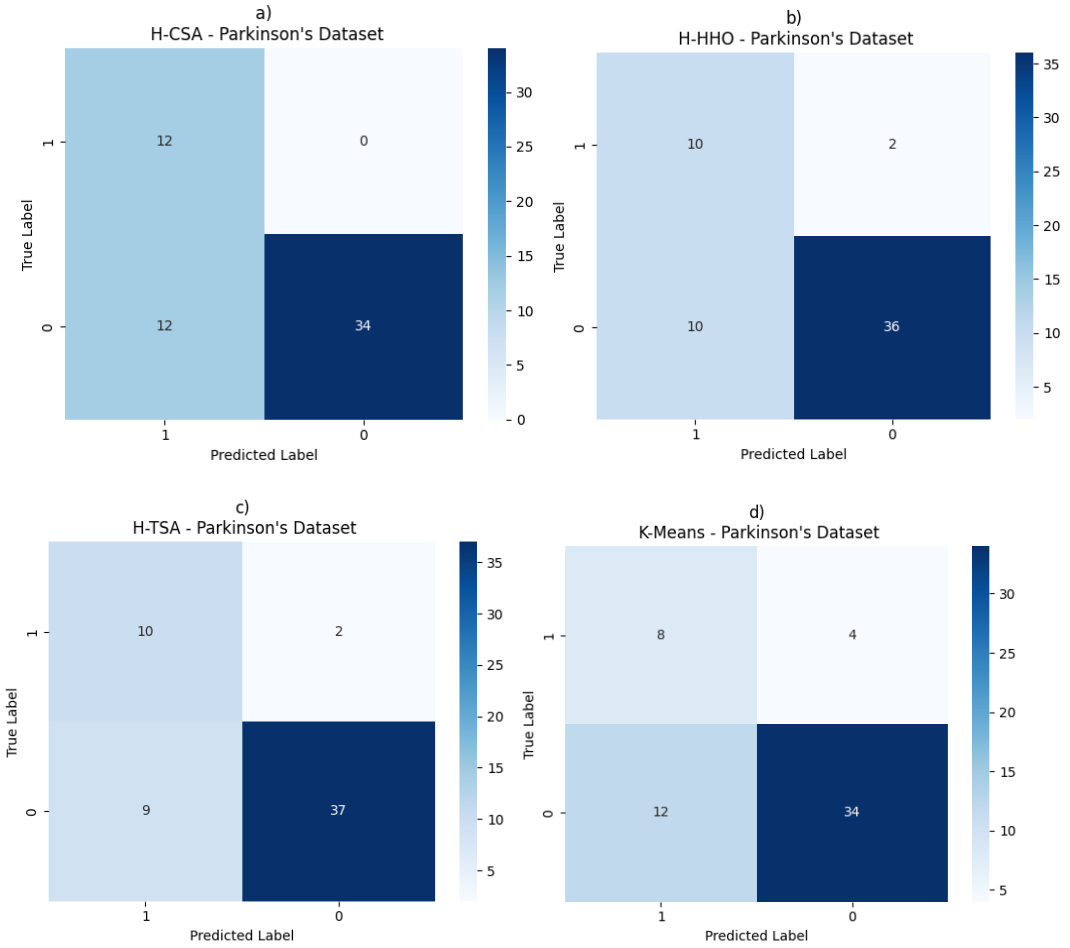

Figure 5. Confusion matrix results of hybrid algorithms and K-Means algorithm for Parkinson's dataset

In Table 9, error metrics results for the Thyroid dataset are shown.

Table 9. Performance results of hybrid algorithms and K-Means algorithm for Thyroid dataset

	H-CSA			H-HHO			H-TSA			K-Means Algorithm		
Algorithm	Train Score: 96.9444% Test Score: 96.2962%			Train Score: 97.8373% Test Score: **96.9907**%			Train Score: 92.0238% Test Score: 90.3703%			Train Score: 43.1151% Test Score: 43.4259%		
Status	**Not Hyperthyroid**	**Hyper Function**	**Subnormal Function**	**Not Hyperthyroid**	**Hyper Function**	**Subnormal Function**	**Not Hyperthyroid**	**Hyper Function**	**Subnormal Function**	**Not Hyperthyroid**	**Hyper Function**	**Subnormal Function**
Accuracy (%)	**0.9847**	0.9652	0.9759	**0.9847**	**0.9722**	**0.9828**	**0.9847**	0.9097	0.9129	0.9171	0.4986	0.4527
Recall (%)	**0.5423**	0.8750	0.9803	**0.5423**	**0.9732**	**0.9824**	**0.5423**	0.3839	0.9436	0.3389	0.6160	0.4268
Specificity (%)	**0.9971**	0.9702	0.9239	**0.9971**	**0.9721**	**0.9883**	**0.9971**	0.9384	0.5555	0.9333	0.4921	0.7543
Precision (%)	**0.8421**	0.6163	0.9933	**0.8421**	**0.6566**	**0.9989**	**0.8421**	0.2544	0.9610	0.1250	0.0622	0.9528
F1-Score (%)	**0.6597**	0.7232	0.9868	**0.6597**	**0.7841**	**0.9906**	**0.6597**	0.3060	0.9523	0.1826	0.1130	0.5895

Types of thyroid disease were examined in the Not Hyperthyroid, Hyperfunction, and Subnormal Function categories. The successful algorithms were shown in green, blue, and red according to the diseases, respectively. H-CSA, H-HHO, and H-TSA algorithms produced the same value with 0.9847 accuracy in classifying not-hyperthyroid disease. The H-HHO algorithm succeeded in diagnosing hyperfunction and Subnormal Function disease types. It achieved high success in all disease types with an accuracy rate exceeding 90% in all three algorithms. For the Thyroid dataset, the H-HHO algorithm with the highest scores in all measurement metrics is relatively more successful than the other two hybrid algorithms. When the H-HHO algorithm is compared with the K-Means algorithm, the accuracy of the H-HHO algorithm is about 7% more successful in classifying Hyperthyroid disease, 97% in classifying Hyper Function disease, and 117% in classifying Subnormal Function disease. In Figure 6, confusion matrices taken using the thyroid dataset are given.

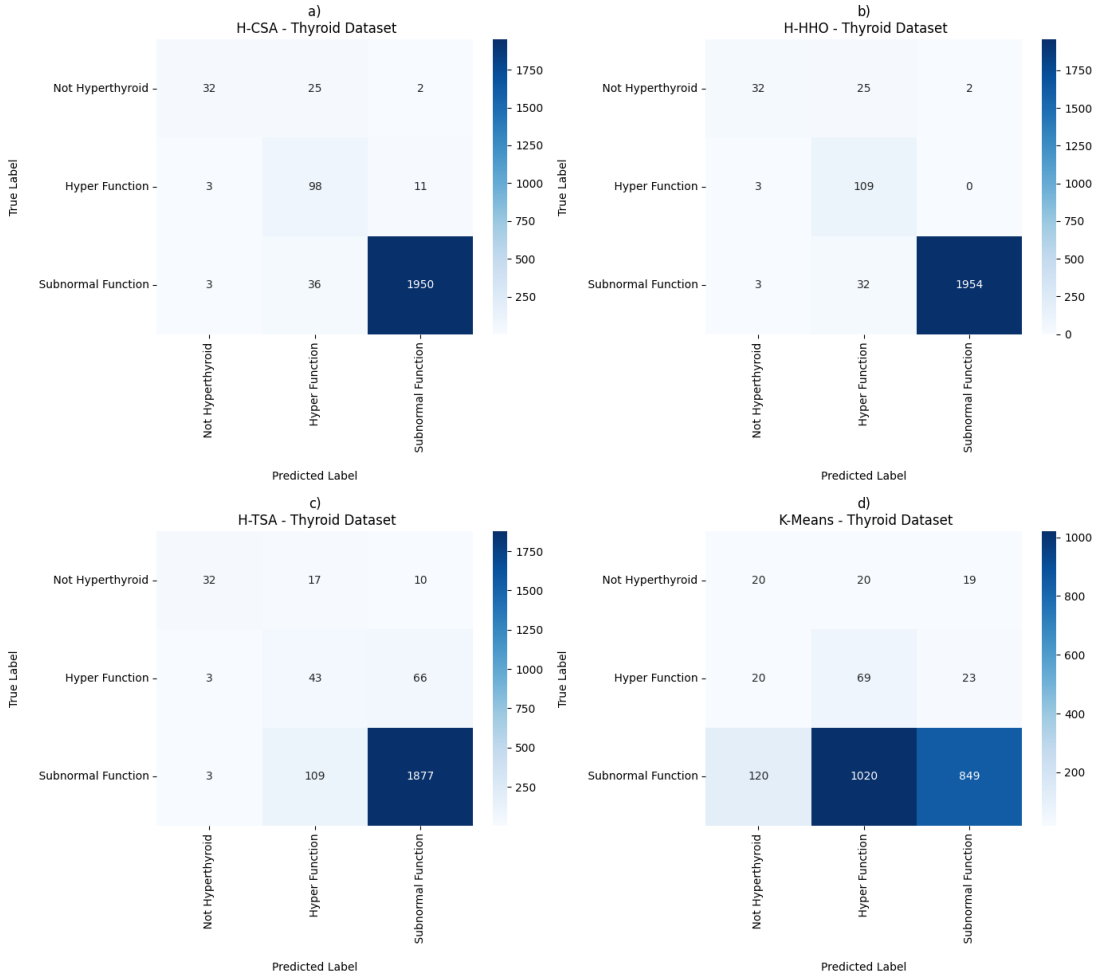

Figure 6. Confusion matrix results of hybrid algorithms and K-Means algorithm for Thyroid dataset

4. Conclusion and Discussion

In the present study, we address the local optimum issue caused by randomly generated initial centroid values in the clustering process of the K-Means algorithm, as well as provide a solution to the challenges in handling large files. We enhance the clustering accuracy by utilizing a metaheuristic algorithm to determine the parameter values for disease diagnosis. In other words, the study aimed to achieve a more precise diagnosis by adjusting the significance level of highly influential parameters. In addition, it sought to prevent false diagnoses by reducing the weight of parameters with minimal impact on the disease diagnosis, with the goal of eliminating false diagnoses by minimizing the effect of irrelevant parameters. By optimizing the coefficients, we aimed to address the issue of the K-Means approach getting trapped in the local optimum and enhance the clustering technique's precision.

The study utilized a hybridized framework that integrated the CSA, TSA, and HHO algorithms and the K-Means method. This approach successfully detected diseases using four distinct medical datasets: dermatology, diabetes, Parkinson's, and thyroid. Additionally, we statistically evaluated the performance measures of the three hybrid algorithms (H-CSA, H-HHO, and H-TSA) and the K-Means algorithm. Among the datasets and optimization algorithms we used in our H-HHO study, the H-HHO algorithm generally achieved higher performance than other algorithms in Diabetes, Dermatology, and Thyroid datasets. On the other hand, the H-CSA achieved the same success rate as H-HHO in the diabetes dataset, although H-HHO produced close results. On the other hand, the H-TSA achieved the highest performance only in the Parkinson's dataset compared to other algorithms. However, H-TSA, in other datasets, is in the last place. We found that the H-CSA algorithm is the fastest of all four data sets used for mean runtime. Despite this, we have concluded that all three algorithms are consistent by producing results that are similar to each other.

The suggested hybrid three metaheuristic algorithm in the Dermatology dataset was detected with 100% success in Psoriasis, Pityriasis Rubra Pilaris, and Chronic Dermatitis disease types. In other disease types in this group (Lichen Planus, Seborrheic Dermatitis, and Pityriasis Rosea), the H-HHO algorithm has been the most successful in the Dermatology dataset, which has six disease types since it has the lowest error value. Test results with the K-Means algorithm demonstrated that the hybrid models were almost half as successful. The results of all three metaheuristic algorithms proposed in the diabetes dataset were determined with the lowest success rate (an average of 73%) compared to the results of other datasets. We predict that this situation is due to the structure of the dataset and the number of input parameters. The H-CSA and H-HHO algorithms produced the same detection value in this dataset and became equivalent algorithms. On the other hand, the K-Means algorithm achieved less success than about 20% of our suggested hybrid algorithms. In the Parkinson's dataset, the H-TSA algorithm, which had a low success rate in other datasets, was the algorithm that produced the most successful diagnosis result. The H-CSA and H-HHO algorithms detected the equal disease in this dataset. In addition, the K-Means algorithm has achieved less success than approximately 12% of the hybrid algorithms we recommend.

The three metaheuristic algorithms proposed in detecting disease Not Hyperthyroid, the subcategory of thyroid disease, achieved the same result with a success rate of 98%. In diagnosing disease types of Hyper Function and Subnormal Function, the H-HHO algorithm achieved higher diagnostic success. Besides, the K-Means algorithm achieved half the diagnostic success of the algorithms we suggested in diseases Hyper Function and Subnormal Function, excluding disease Not Hyperthyroid. The presented study offers significant advantages such as test accuracy rate, staff workload, test costs, and waiting time for the test results detection of different types of diseases in medical datasets. In future research, new studies can be carried out using datasets of different diseases in the health field and different metaheuristic algorithms or deep learning methods.

Conflict of Interest Statement

The authors declare that there is no conflict of interest

References

[1] I. Zheliznyak, Z. Rybchak, and I. Zavuschak, "Analysis of clustering algorithms.," in *Advances in Intelligent Systems and Computing: Selected Papers from the International Conference on Computer Science and Information Technologies, September 6-10 , 2017. Lviv, Ukraine.* Springer, 2017, pp. 305–314.

[2] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm.," *Applied soft computing*, vol. 11, no. 1, pp. 652–657, 2011. doi.org/10.1016/j.asoc.2009.12.025

[3] M. Ren, J. Zhang, L. Khoukhi, H. Labiod, and V. Vèque, "A unified framework of clustering approach in vehicular ad hoc networks.," *IEEE Transactions on intelligent transportation systems*, vol. 19, no. 5, pp. 1401–1414, 2017. doi.org/10.1109/TITS.2017.2727226

[4] M.M. Rahman and D.N. Davis, "Addressing the class imbalance problem in medical datasets.," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, p. 224, 2013. doi.org/10.7763/IJMLC.2013.V3.307

[5] S. Ding, H. Jia, M. Du, and Y. Xue, "A semi-supervised approximate spectral clustering algorithm based on HMRF model," *Information Sciences*, vol. 429, pp. 215–228, 2018. doi.org/10.1016/j.ins.2017.11.016

[6] M.W. Berry, A. Mohamed, and B.W. Yap, "Supervised and unsupervised learning for data science," Springer, 2019.

[7] C. Fraley and A.E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The computer journal*, vol. 41, no. 8, pp. 578–588, 1998. doi.org/10.1093/comjnl/41.8.578

[8] A. Saxena, M. Prasad, A. Gupta, et al., "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017. doi.org/10.1016/j.neucom.2017.06.053

[9] T. Yi, M. Shi, and H. Zhu, "Medical data publishing based on average distribution and clustering," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 3, pp. 381–394, 2022. doi.org/10.1049/cit2.12094

[10] J. Si, Z. Tian, D. Li, et al., "A multi-modal clustering method for traditonal Chinese medicine clinical data via media convergence," *CAAI Transactions on Intelligence Technology*, p. 2023. doi.org/10.1049/cit2.12230

[11] S. Dörterler, "KANSER HASTALIĞI TEŞHİSİNDE ÖLÜM OYUNU OPTİMİZASYON ALGORİTMASININ ETKİS," *Mühendislik Alanında Uluslararası Araştırmalar VIII*, p. 15, 2023.

[12] S. Dörterler, H. Dumlu, D. Özdemir, and H. Temurtaş, "Hybridization of K-means and Meta-Heuristics Algorithms for Heart Disease Diagnosis," *in NEW TRENDS IN ENGINEERING AND APPLIED NATURAL SCIENCES*, pp. 55. Duvar Publishing (2022).

[13] S. Dörterler, H. Dumlu, D. Özdemir, and H. Temurtaş, "Melezlenmiş K-means ve Diferansiyel Gelişim Algoritmaları ile Kalp Hastalığının Teşhisi," in *International Conference on Engineering and Applied Natural Sciences*, May 10-12, 2022. Konya, Turkey, ICEANS 2022. pp. 1840–1844.

[14] A. Tekerek and M. Dörterler, "The adaptation of gray wolf optimizer to data clustering," *Politeknik Dergisi*, p. 1, 2020. doi.org/10.2339/politeknik.778630

[15] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7425–7440, 2018. doi.org/10.1109/TWC.2018.2867180

[16] F. Hu, H. Chen, and X. Wang, "An Intuitionistic Kernel-Based Fuzzy C-Means Clustering Algorithm With Local Information for Power Equipment Image Segmentation," *IEEE Access*, vol. 8, pp. 4500–4514, 2020. doi.org/10.1109/ACCESS.2019.2963444

[17] J. Maia, C.A.S. Junior, F.G. Guimarães, et al., "Evolving clustering algorithm based on mixture of typicalities for stream data mining," *Future Generation Computer Systems*, vol. 106, pp. 672–684, 2020. doi.org/10.1016/j.future.2020.01.017

[18] R.M. Alguliyev, R.M. Aliguliyev, and L. V Sukhostat, "Efficient algorithm for big data clustering on single machine," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 9–14, 2020. doi.org/10.1049/trit.2019.0048

[19] Y. Li, J. Qi, X. Chu, and W. Mu, "Customer Segmentation Using K-Means Clustering and the Hybrid Particle Swarm Optimization Algorithm," *The Computer Journal,* p. 2022. doi.org/10.1093/comjnl/bxab206

[20] S.H. Shihab, S. Afroge, and S.Z. Mishu, "RFM based market segmentation approach using advanced k-means and agglomerative clustering: a comparative study," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). February 7-9, 2019*. IEEE , pp. 1–4.

[21] A. Fallah Tehrani and D. Ahrens, "Modified sequential k-means clustering by utilizing response: A case study for fashion products," *Expert Systems,* vol. 34, no. 6, p. e12226, 2017. doi.org/10.1111/exsy.12226

[22] O. Akman, T. Comar, D. Hrozencik, and J. Gonzales, "Data clustering and self-organizing maps in biology," in *Algebraic and Combinatorial Computational Biology*, pp. 351–374. Elsevier (2019). doi.org/10.1016/B978-0-12-814066-6.00011-8

[23] J.P.L. Johnson, "Clustering statistics, roughness feedbacks, and randomness in experimental step-pool morphodynamics," *Geophysical Research Letters*, vol. 44, no. 8, pp. 3653–3662, 2017. doi.org/10.1002/2016GL072246

[24] L. Wen, K. Zhou, and S. Yang, "A shape-based clustering method for pattern recognition of residential electricity consumption," *Journal of cleaner production*, vol. 212, pp. 475–488, 2019. doi.org/10.1016/j.jclepro.2018.12.067

[25] K.P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020. doi.org/10.1109/ACCESS.2020.2988796

[26] R.-Q. Liu, Y.-C. Lee, and H.-L. Mu, "Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company," *Knowledge Management Research*, vol. 19, no. 4, pp. 59–76, 2018. doi.org/10.15813/kmr.2018.19.4.004

[27] M. Mehrazma and B. Mahboobi, "Customer Clustering by Combining the Particle Swarm And K-Means Algorithms and Analyzing Their Behavior on Commercial Websites," *Signal Processing and Renewable Energy*, vol. 5, no. 2, pp. 83–97, 2021. dorl.net/dor/20.1001.1.25887327.2021.5.2.6.7

[28] Y. Tang, Y. Guo, Q. Sun, et al., "Industrial polymers classification using laser-induced breakdown spectroscopy combined with self-organizing maps and K-means algorithm," *Optik*, vol. 165, pp. 179–185, 2018. doi.org/10.1016/j.ijleo.2018.03.121

[29] J. Fonseca, G. Douzas, and F. Bacao, "Improving Imbalanced Land Cover Classification with K-Means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures," *Information,* vol. 12, no. 7, p. 266, 2021. doi.org/10.3390/info12070266

[30] M.R. Kanthan and S.N. Sujatha, "Rain drop detection and removal using k-means clustering," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). pp. 1–5. IEEE ,10-12 December 2015, Madurai, India*. doi.org/10.1109/ICCIC.2015.7435707

[31] D. Yu, X. Zhou, Y. Pan, Z. Niu, and H. Sun, "Application of Statistical K-Means Algorithm for University Academic Evaluation," *Entropy,* vol. 24, no. 7, p. 1004, 2022. doi.org/10.3390%2Fe24071004

[32] J. Heil, V. Häring, B. Marschner, and B. Stumpe, "Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils." *Geoderma*, vol. 337, pp. 11–21, 2019. doi.org/10.1016/j.geoderma.2018.09.004

[33] N. Xiaoying, X. Liya, and Z. Xiaoyu, "Classification of Rice According to the Geographic Origin Based on K-Means Cluster and Hierarchinal Cluster [J]," *Journal of Agricultural Mechanization Research,* vol. 6, p. 2012.

[34] T.K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis & Prevention,* vol. 41, no. 3, pp. 359–364, 2009. doi.org/10.1016/j.aap.2008.12.014

[35] F. Bação, V. Lobo, and M. Painho, "Self-organizing maps as substitutes for k-means clustering," in *International Conference on Computational Science, 22-25 May 2005,* pp. 476–483. Springer, Atlanta, USA. doi.org/10.1007/11428862_65

[36] S. Chakraborty, N.K. Nagwani, and L. Dey, "Weather forecasting using incremental K-means clustering," *arXiv preprint arXiv:1406.4756*. p. 2014. doi.org/10.48550/arXiv.1406.4756

[37] D.E. Kurniawan and A. Fatulloh, "Clustering of Social Conditions in Batam, Indonesia Using K-Means Algorithm and Geographic Information System," *International Journal of Earth Sciences and Engineering (IJEE),* vol. 10, no. 05, pp. 1076–1080, 2017. doi.org/10.31227/osf.io/qc342

[38] J. Zhou and L.C. Le Liang, "Geographic profiling based on multi-point centrography with K-means clustering," p. 2012. doi.org/10.5281/zenodo.1058785

[39] R.S. Evans, J.F. Lloyd, G.J. Stoddard, J.R. Nebeker, and M.H. Samore, "Risk factors for adverse drug events: a 10-year analysis," *Annals of Pharmacotherapy*, vol. 39, no. 7–8, pp. 1161–1168, 2005. doi.org/10.1345/aph.1e642

[40] E.R. Vickers, H. Boocock, R.D. Harris, et al., "Analysis of the acute postoperative pain experience following oral surgery: identification of 'unaffected','disabled'and 'depressed, anxious and disabled'patient clusters," *Australian dental journal*, vol. 51, no. 1, pp. 69–77, 2006. doi.org/10.1111/j.1834-7819.2006.tb00404.x

[41] D.M. Falkenbach, E.E. Reinhard, and M. Zappala, "Identifying psychopathy subtypes using a broader model of personality: An investigation of the five factor model using model-based cluster analysis," *Journal of interpersonal violence*, vol. 36, no. 15–16, pp. 7161–7184, 2021. doi.org/10.1177/0886260519831388

[42] H. Kim, B. Kim, S.H. Kim, C.H.K. Park, E.Y. Kim, and Y.M. Ahn, "Classification of attempted suicide by cluster analysis: A study of 888 suicide attempters presenting to the emergency department," *Journal of affective disorders*, vol. 235, pp. 184–190, 2018. doi.org/10.1016/j.jad.2018.04.001

[43] R. Shai, T. Shi, T.J. Kremen, et al., "Gene expression profiling identifies molecular subtypes of gliomas," *Oncogene*, vol. 22, no. 31, pp. 4918–4923, 2003. doi.org/10.1038/sj.onc.1206753

[44] F. Bertucci, P. Finetti, J. Rougemont, et al., "Gene expression profiling identifies molecular subtypes of inflammatory breast cancer," *Cancer research*, vol. 65, no. 6, pp. 2170–2178, 2005. doi.org/10.1158/0008-5472.can-04-4115

[45] K. Ushizawa, C.B. Herath, K. Kaneyama, et al., "cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period," *Reproductive Biology and Endocrinology*, vol. 2, no. 1, pp. 1–16, 2004. doi.org/10.1186/1477-7827-2-77

[46] E. Şahin, D. Özdemir, and H. Temurtaş, "Multi-objective optimization of ViT architecture for efficient brain tumor classification," *Biomedical Signal Processing and Control*, vol. 91, p. 105938, 2024. doi.org/10.1016/j.bspc.2023.105938

[47] N. Yagmur, İ. Dag, and H. Temurtas, "Classification of anemia using Harris hawks optimization method and multivariate adaptive regression spline," *Neural Computing and Applications*, pp. 1–20, 2024. doi.org/10.1007/s00521-023-09379-y

[48] N.N. Arslan, D. Ozdemir, and H. Temurtas, "ECG heartbeats classification with dilated convolutional autoencoder," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 417–426, 2024. doi.org/10.1007/s11760-023-02737-2

[49] N. Yagmur, I. Dag, and H. Temurtas, "A new computer-aided diagnostic method for classifying anaemia disease: Hybrid use of Tree Bagger and metaheuristics," *Expert Systems*, p. e13528, 2023. doi.org/10.1111/exsy.13528

[50] F. Aydemir and S. Arslan, "Covid-19 pandemi sürecinde çocukların el yıkama alışkanlığının nesnelerin interneti tabanlı sistem ile izlenmesi," *Mühendislik Bilimleri ve Araştırmaları Dergisi,* vol. 3, no. 2, pp. 161–168, 2021. doi.org/10.46387/bjesr.949311

[51] N. Yağmur, H. Temurtaş, and İ. Dağ, "Anemi hastalığının yapay sinir ağları yöntemleri kullanılarak sınıflandırılması," *Journal of Scientific Reports-B,* no. 008, pp. 20–34, 2023.

[52] A.M. Ikotun, M.S. Almutari, and A.E. Ezugwu, "K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions," *Applied Sciences,* vol. 11, no. 23, p. 11246, 2021. doi.org/10.3390/app112311246

[53] M. Delattre and P. Hansen, "Bicriterion cluster analysis.," IEEE Transactions on Pattern Analysis and Machine Intelligence. no. 4, pp. 277–291, 1980. https://doi.org/10.1109/TPAMI.1980.4767027

[54] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR),* vol. 31, no. 3, pp. 264–323, 1999. doi.org/10.1145/331499.331504

[55] A.G.D.Ç. TAŞKIN and G.G. Emel, "Veri Madenciliğinde Kümeleme Yaklaşimlari Ve Kohonen Ağlari İle Perakendecilik Sektöründe Bir Uygulama," *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* vol. 15, no. 3, pp. 395–409, 2010.

[56] L. Zhang, M. Jin, and S.-J. Yoo, "A clustering detector with graph theory for blind detection of spatial modulation systems," *Wireless Networks,* vol. 27, no. 2, pp. 1193–1201, 2021. doi.org/10.1007/s11276-020-02508-8

[57] M. Sharma, G.N. Purohit, and S. Mukherjee, "Information retrieves from brain MRI images for tumor detection using hybrid technique K-means and artificial neural network (KMANN)," in *Networking communication and data knowledge engineering*, pp. 145–157, Springer (2018). doi.org/10.1007/978-981-10-4600-1_14

[58] J. Estupiñán Ricardo, J.J. Domínguez Menéndez, I.F. Barcos Arias, J.M. Macías Bermúdez, and N. Moreno Lemus, "Neutrosophic K-means for the analysis of earthquake data in Ecuador," *Neutrosophic Sets and Systems,* vol. 44, no. 1, p. 29, 2021.

[59] M. Roberto e Souza, A. Carlos Sousa e Santos, and H. Pedrini, "A Hybrid Approach Using the k-means and Genetic Algorithms for Image Color Quantization," *Recent Advances in Hybrid Metaheuristics for Data Clustering*, pp. 151–171, 2020. doi.org/10.1002/9781119551621.ch9

[60] S. Kumar, V.K. Solanki, S.K. Choudhary, A. Selamat, and R. González Crespo, "Comparative Study on Ant Colony Optimization (ACO) and K-Means Clustering Approaches for Jobs Scheduling and Energy Optimization Model in Internet of Things (IoT)," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 1, p. 2020. doi.org/10.9781/ijimai.2020.01.003

[61] A.K. Abasi, A.T. Khader, M.A. Al-Betar, S. Naim, Z.A.A. Alyasseri, and S.N. Makhadmeh, "A novel hybrid multi-verse optimizer with K-means for text documents clustering," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17703–17729, 2020. doi.org/10.1007/s00521-020-04945-0

[62] A. Kaur, S.K. Pal, and A.P. Singh, "Hybridization of Chaos and Flower Pollination Algorithm over K-Means for data clustering," *Applied Soft Computing*, vol. 97, p. 105523, 2020. doi.org/10.1016/j.asoc.2019.105523

[63] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection," *Knowledge-Based Systems*, vol. 203, p. 106167, 2020. doi.org/10.1016/j.knosys.2020.106167

[64] S. Aggarwal and P. Singh, "Cuckoo and krill herd-based k-means++ hybrid algorithms for clustering," *Expert Systems,* vol. 36, no. 4, p. e12353, 2019. doi.org/10.1111/exsy.12353

[65] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham, *The UCI Machine Learning Repository*. USA, 2022.

[66] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, 2020. doi.org/10.1007%2Fs40200-020-00520-5

[67] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *Nature Precedings*, p. 1, 2008. doi.org/10.1109%2FTBME.2008.2005954

[68] M. Hosseinzadeh, O.H. Ahmed, M.Y. Ghafour, et al., "A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3616–3637, 2021. doi.org/10.1007/s11227-020-03404-w

[69] H.A. Güvenir, G. Demiröz, and N. Ilter, "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals," *Artificial intelligence in medicine*, vol. 13, no. 3, pp. 147–165, 1998. doi.org/10.1016/s0933-3657(98)00028-1

[70] M. Ahmed, R. Seraj, and S.M.S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics,* vol. 9, no. 8, p. 1295, 2020. doi.org/10.3390/electronics9081295

[71] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future generation computer systems*, vol. 97, pp. 849–872, 2019. doi.org/10.1016/j.future.2019.02.028

[72] M.S. Kiran, "TSA: Tree-seed algorithm for continuous optimization," *Expert Systems with Applications*, vol. 42, no. 19, pp. 6686–6698, 2015. doi.org/10.1016/j.eswa.2015.04.055

[73] M. Aslan, M. Beskirli, H. Kodaz, and M.S. Kiran, "An improved tree seed algorithm for optimization problems," *Int J Mach Learn Comput*, vol. 8, no. 1, pp. 20–25, 2018. doi.org/10.18178/ijmlc.2018.8.1.657

[74] A. Beşkirli, D. Özdemir, and H. Temurtaş, "A comparison of modified tree–seed algorithm for high-dimensional numerical functions," *Neural Computing and Applications*, vol. 32, no. 11, pp. 6877–6911, 2020. doi.org/10.1007/s00521-019-04155-3

[75] A. Askarzadeh, "A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm," *Computers & Structures,* vol. 169, pp. 1–12, 2016. doi.org/10.1016/j.compstruc.2016.03.001

[76] Y. Li and H. Wu, "A clustering method based on K-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012. doi.org/10.1016/j.phpro.2012.03.206

[77] J.M. Pena, J.A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm," *Pattern recognition letters,* vol. 20, no. 10, pp. 1027–1040, 1999. doi.org/10.1016/S0167-8655(99)00069-0

[78] A. Ucan, M. Dörterler, and E. Akçapınar Sezer, "An emotion analysis scheme based on Gray Wolf optimization and deep learning," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 13, p. e6204, 2021. doi.org/10.1002/cpe.6204

[79] H. Karamollaoğlu, İ.A. Doğru, M. Dörterler, A. Utku, and O. Yıldız, "Sentiment analysis on Turkish social media shares through lexicon based approach," in *2018 3rd International Conference on Computer Science and Engineering (UBMK). September 20-23, 2018*. IEEE, pp. 45–49.

[80] F. Aydemir and S. Arslan, "A System Design With Deep Learning and IoT to Ensure Education Continuity for Post-COVID," *IEEE Transactions on Consumer Electronics*, p. 2023. doi.org/10.1109/TCE.2023.3245129