Research Article

Optimized grass hopper algorithm for diagnosis of Parkinson's disease

Shallu Sehgal[1] · Manisha Agarwal[1] · Deepak Gupta[2] · Shirsh Sundaram[2] · Arun Bashambu[1]

Received: 31 October 2019 / Accepted: 25 April 2020

Abstract
The Modified Grasshopper Optimization Algorithm which identifies Parkinson disease symptoms at an early (premature) stage was proposed. Parkinson disease, type of movement ailment, could be life-threatening if not treated at premature stage. Therefore, diagnosis of Parkinson disease became essential in early stages so that all the symptoms could be controlled by giving required medication to the patient. Hence ensuring the patient longevity. As part of this research work, a novel model Modified Grasshopper Optimization Algorithm was introduced which was based on the traditional Grasshopper Optimization Algorithm and search strategy for feature selection. Grasshopper Optimization Algorithm was relatively a novel heuristic optimization swarm intelligence algorithm which was stimulated by grasshoppers searching for food. This population-based method has capability to provide solution for real-life problems in undefined search space. It mimics grasshopper swarm's behaviour and their social interaction. Popular algorithms like Random Forest, Decision Tree and k-Nearest Neighbour classifier were used in judgement on shortlisted aka selected features. Different datasets of handwriting (meander and spiral), speech and voice were used for evaluating the presented model. The proposed algorithm was effective in Parkinson disease identification having accuracy (computed) of 95.37%, 99.47% detection rate and 15.78% false alarm rate. This helps larger cause of patient in receiving treatment in pre-mature stage. The presented bio-inspired algorithm was adequately steady and has ability to identify the optimal feature set. Finally results obtained from the assessment of introduced Modified Grasshopper Optimization Algorithm on these data sets were evaluated and contrasted with respect to outcome of Modified Grey Wolf Optimizer and Optimized Cuttlefish Algorithm. The experiment's outcome revealed that the presented Modified Grasshopper Optimization Algorithm assists in reducing the selected features count and improving the accuracy.

Keywords Modified grasshopper optimization algorithm · Parkinson's disease · Feature selection · Machine learning

1 Introduction

Parkinson's disease, a disorder of the nervous system, happens because of the loss of brain cells. It impacts body mobility. Its symptoms slowly become evident. A few of these symptoms which appear at the initial stage are slowness in movement, tremors, rigidness in muscles, poor body posture, imbalance, deviation in speech and handwriting strokes. In this disease, patient's nerve cells slowly lose the capacity of communicating among themselves that leads to nervous system disorders like depression etc. This disorder should be detected in its initial stage as it's incurable. If exact symptoms of Parkinson's disease are known with their relative weightage then medical practitioners can recommend pathology lab tests for those features and disease detection can happen on the first consultation itself. This will lead to early detection of Parkinson disease. Symptoms like change in handwriting strokes and speaking patterns help in early-stage detection of this disease. Erdogu Sakar & team recently collected

✉ Deepak Gupta, deepakgupta@mait.ac.in; Shallu Sehgal, shallus@rediffmail.com; Manisha Agarwal, manishaagarwal18@yahoo.co.in; Shirsh Sundaram, sundaramshirsh@gmail.com; Arun Bashambu, arunbashambu@rediffmail.com | [1]Banasthali Vidyapeeth, Banasthali, Jaipur, India. [2]Maharaja Agrasen Institute of Technology, Rohini, Delhi, India.

Published online: 02 May 2020

a speech data set by inspecting the pronunciation of vowels 'a' and 'o' [1] of disease impacted people. Other than speaking pattern, handwriting strokes pattern may assist in identifying the ailment. Handwriting analysis can be done by studying and comparing the figures created by the ailing person and fit person. In this test, an individual is directed to create figures of spiral and meanders in air. The individual is asked to draw it in black and white in both clockwise and anticlockwise directions. Factors reviewed for differentiating a patient from a healthy person are fare handedness(right/left), individual age, mean and maximum distance between provided outline in test, handwriting strokes recorded in the test and drawing time duration. Given dataset in [2] is obtained from the latest research endeavour by University of São Paulo State (Botucatu Medical School)—Brazil, where patients statistics were gathered and compiled together for analysis. In this study, we have deployed an optimized Grasshopper algorithm to find out if the person is suffering from disease or not based on collected voice sample data in [1–3]. A systematic study and analysis of the technologies, tools for the treatment of this disease has been carried out [4]. Since the last two decades, data has increased multi-fold in terms of numbers of features and instances. Due to a huge increase in data volume, many data-related problems creep in and noise increases. Increased instance count and feature count leads to a lot of pre-processing and processing work. Increased noise causes performance drop and results get degraded. Thus, it is necessary to do data treatment. A high volume of data leads to an increase in computation cost and complexity. Feature Selection comes handy in keeping the data volume in control, hence reduces complexity and computation cost. It avoids model over-fitting and evades curse of dimensionality. It plays a vital role in building ML models. Feature selection is also called variable or attribute selection. In this, a subset of features is selected from all available features. The main aspect that is kept in consideration is model accuracy, which is computed before and after feature selection. Feature selection has been classified into a filter-based and wrapper-based algorithm. The filter-based algorithm makes use of statistical methodology for finding the importance of each feature (attribute). Wrapper-based algorithm computation makes use of the Machine Learning algorithm approach. Wrapper-based algorithm computation is expensive than a filter-based algorithm. Wrapper methods are further categorized as Sequential Search Algorithm and Heuristic Search Algorithm. Nowadays, effective feature extraction is essential for model building and is used to categorize bigdata into social IoT [5]. IoT applications have been detailed in [6, 7].

The objective is to develop a more optimized & efficient algorithm, including the following:

- To reduce the noise in data by using a feature selection method, in turn decreasing the time complexity.
- To increase the algorithm accuracy by developing an optimized feature selection machine learning algorithm for Parkinson disease.

Evolutionary algorithms (subset of Artificial Intelligence) have focus on Biological evolution. Biological evolution processes are reproduction, mutation, recombination and selection. Evolutionary algorithm is based on random sampling, which is different from previous optimization techniques. For the solution, the process of biological evolution is repeatedly applied on the given population. The fitness function determines the solution quality. The solution changes as per the evolutionary process, finally getting a universal solution to the problem. Evolutionary algorithms don't consider fundamental fitness landscape and these are performing well under diverse circumstances. A basic evolutionary algorithm is used to solve the complex problem but its computation cost is very high (setback) [8]. Fitness function approximation reduces this setback. Evolutionary algorithms can be modified for usage in feature selection viz. Binary Bat Algorithm [9], Modified Binary Bat Algorithm [10], and Binary chaotic crow search algorithm [11].

Research work in Meta-heuristic and evolutionary algorithms has increased a lot recently. Base of these algorithms are genetic and nature-inspired algorithms. These algorithms compute fitness function, which does the task of eliminating the inconsequential solutions and then focuses on the significant solutions leading to optimization of the genetic lineage (uses mutation prevalent theories and survival of the fittest). Genetic methodologies, evolutionary theories and evolutionary methodologies are different evolutionary algorithms [8]. Xin-She Yang proposed a nature-inspired Firefly Algorithm, which uses fireflies flashing attributes [12]. Bat Algorithm [13] has been proposed by him getting influenced by echoing features of bats. Ant colony optimization algorithm [14] and gravitational search algorithm use correlation amid gravity and mass [5]. Pham and associates proposed the Bees Algorithm [15]. Erik Cuevas and team presented CAB algorithm [16]. The Group Grey Wolf Optimization (GGWO) methodology is used to improve the Alzheimer disease detection performance using DT, KNN, and CNN classifiers [17]. Three optimized bioinspired algorithms, Crow Search Algorithm (ICSA), Grey Wolf Algorithm (IGWA) and Cuttlefish Algorithm (ICFA) are used for feature selection from images in detecting prevalent lung disorders Chronic Obstructive Pulmonary disease and fibrosis [18].

For doing feature selection of software-based usability models [19], many bio-inspired algorithms had been developed viz. modified crow search algorithm [20], modified binary bat algorithm [10] and modified whale optimization algorithm [21]. Using grey wolf optimizer [22], optimal features of Thyroid disease get identified. This enables synchronization [23] and unique identification [24] of the predicted disease.

Using the above ideas, we have presented a modified Grasshopper Optimization Algorithm (GOA) which is inspired by grasshoppers searching for food. Grasshoppers explore the available search region by using repulsive forces from other grasshoppers and then they find promising areas by using attractive forces from other grasshoppers. Modified Grasshopper Optimization Algorithm (MGOA) is an extended and optimized form of original GOA. GOA has the capability of exploring all agents for getting to the next position and helps in solving hidden search space problems [25]. GOA is used for optimizing machine learning factors and choosing features. It's the ability to solve real-life problems in undefined search space has been proved. We have proposed an optimized algorithm which uses different fitness functions. MGOA has been implemented on Parkinson's speech and HandPD datasets for diagnosis of Parkinson's disease. High accuracy results lead to reduction in disease detection time and hence lead to early detection and treatment. MGOA provides a small selected feature set having very high accuracy which can be used by practitioners for early accurate detection of Parkinson disease. The algorithm efficiently gave an average accuracy of 95.37%, this algorithm is stimulated by the biological behaviour found in swarms of grasshoppers.

In this paper [26] a novel optimization scheme, namely Chronological-GOA for genetic data filtering and classification of cancer has been presented. In this paper [27] GOA and SVM approach has been applied on bio-medical records of Iraq's cancer patients in year 2010–2012 and also for California University Irvine data sets.

The main contributions of this work are:

- New swarm-based bio-inspired Modified Grasshopper Optimization Algorithm (MGOA) for finding the optimal subset of feature has been proposed.
- MGOA is based on traditional Grasshopper Optimization Algorithm and is evaluated on voice PD, sound recordings and Hand PD (both spiral and meander); for the diagnosis of Parkinson ailment having improved accuracy 95.37%
- Three ML algorithms have been applied for correctly classifying the above datasets. These algorithms are Random Forest, K-Nearest Neighbour and Decision Tree.
- The performance of the model is calculated using the confusion metrics. Then accuracy, Detection Rate and False Alarm Rate are obtained to assess the algorithm.
- Results show that the Random Forest algorithm applied in MGOA beats other machine learning models.

This paper constitutes of five sections. Section 2 contains the background discussing about traditional grasshopper optimization algorithm. The presented algorithm *MGOA* for selecting feature and its implementation is shown in below Sect. 3. In Sect. 4, comparisons and results of MGOA with OCFA and MGWO are detailed. In Sect. 5, conclusion and possible extensions for future work are discussed.

2 Background

2.1 Traditional GOA

In 2017, Saremi S, Mirjalili S, Lewis Andrew proposed a Grasshopper Optimization Algorithm. Grasshopper are a type of insect. These pests are main cause which lead to damage of crops and hence reducing the agriculture yield. Grasshoppers generally live independently individually in nature; however, they are known to form largest swarm of all creatures when they join in [28].

Grasshoppers life cycle as shown in above Fig. 1a has three stages viz. Egg, Nymph and Adult. Grasshopper's uniqueness is that it displays swarming tendency in

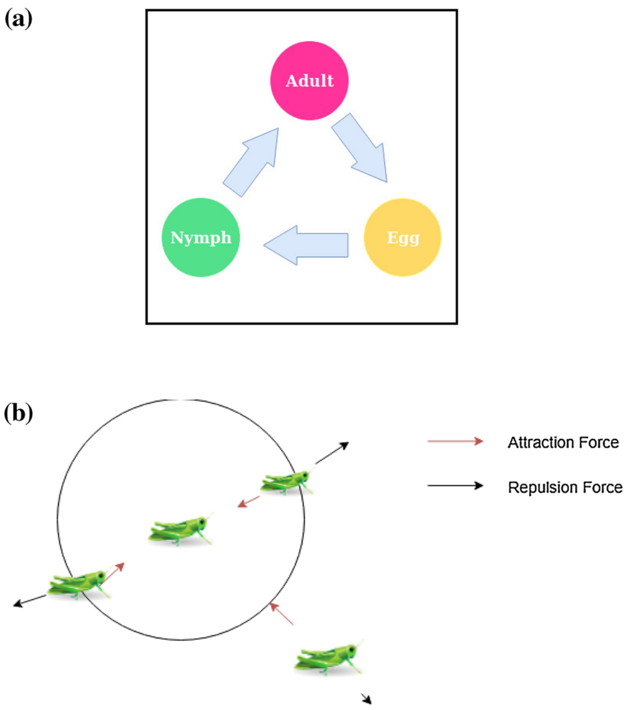


Fig. 1 **a** Life cycle of Grasshopper, **b** Grasshopper swarm's corrective primitive patterns

nymph stage and adulthood [29]. Millions of Grasshoppers in nymph stage jump together and their movement is like rolling cylinders. These nymph grasshoppers eat almost all vegetation in their path. On becoming adult, these grasshoppers form an aerial swarm. In this way these grasshoppers actually travel and spread over very large area. In the larval stage, the chief characteristics are their sluggish movement and mini steps. In contrast to this, adult swarm's features are long- range and abrupt movement. Grasshoppers swarms have another characteristic viz. food source seeking.

In nature inspired algorithms like GOA, search has exploration and exploitation phases. The search agents traverse suddenly and unexpectedly in exploration phase, where as they move locally in exploitation phase. Grasshoppers naturally perform these two functions and target seeking. By representing this behaviour in a new mathematical model, a new nature-inspired algorithm gets designed.

Grasshoppers swarming behaviour is represented by the mathematical model in [30]:

$$\mathrm{Xi} = \mathrm{Si} + \mathrm{Gi} + \mathrm{Ai} \tag{1}$$

In this mathematical model, ith position of Grasshopper's is $\mathrm{X_i}$. $\mathrm{S_i}$ is social interplay and gravitational force due on ith grasshopper is Gi. Air advection is represented by Ai.

Randomness in behaviour is given by Eq. 2 below

$$X_i = r_1 S_i + r_2 G_i + r_3 A_i \tag{2}$$

here r_1, r_2, and r_3 are random number between 0 and 1.

The social interplay's computation is

$$S_i = \sum_{\substack{j=1 \\ j!=i}}^{N} s\left(d_{ij}\right)\widehat{d_{ij}} \tag{3}$$

Here the distance between ith and jth grasshopper is represented by d_{ij}, computed as.

$d_{ij} = \left|x_j - x_i\right|$. Social force's strength is given by function s in Eq. (4). d_{ij} unit vector between the ith grasshopper and jth grasshopper is computed as $d_{ij} = \frac{x_j - x_i}{d_{ij}}$.

The social force represented by function s is computed as below:

$$S(\Gamma) = f\mathrm{e}^{\frac{-\Gamma}{l}} - \mathrm{e}^{-\Gamma} \tag{4}$$

here f is attraction intensity, l represents the attractive length scale.

This S function indicates effect of attractive and repulsive social forces of grasshopper. Consideration range of Distance is from 0 to 15. Repulsion spread is from 0 to 2.079. Comfortable displacement between grasshoppers is 2.079 units, as attraction or repulsion is not there for grasshoppers when they are not in range of 2.079 units. It's known as comfortable zone. In case of artificial grasshoppers, it has been observed that there is social behaviour difference with the change in l and f parameter as shown by Eq. (4). Parameters l and f vary independently of each other, their effect is observed on function s.

Attraction, repulsion region and comfort zone get changed proportionately by changing the parameters l and f. For values like (l = 1.0 or f = 1.0), attraction or repulsion area tend to be very small. l = 1.5 and f = 0.5 are selected from the range [25].

Functions in Fig. 1b illustrates concept of comfort zone and grasshopper interaction's conceptual model. In simple form, motivating force found was this social interaction in earlier locust swarming models [26]. Function s helps in dividing the region into comfort zone, attraction region and repulsion region between two grasshoppers. Value close to zero gets returned for distances greater than 10. This function can't be used for the large distance between the grasshoppers as the strong forces can't be applied. This problem is overcome by keeping (and mapping) the distance of grasshoppers in the range between 1 and 4. In Eq. (1)

G component is computed as:

$$G_i = -g\overrightarrow{e_g} \tag{5}$$

here gravitational constant is g and unity vector towards the earth's centre is $\overrightarrow{e_g}$. Also, A component referred in Eq. (1) is computed

$$A_i = u\overrightarrow{e_w} \tag{6}$$

here u refers constant drift, unity vector towards wind's direction is $\overrightarrow{e_w}$. The nymph grasshoppers are wingless and nymph movement is highly correlated with direction of wind. On substitution of S, A and G in Eq. (1), the equation gets expanded as:

$$X_i = \sum_{\substack{j=1 \\ j!=i}}^{N} s\left(\left|x_j - x_i\right|\right)\frac{x_j - x_i}{d_{ij}} - d\widehat{e_g} + u\widehat{e_w} \tag{7}$$

where $S(\Gamma) = f\mathrm{e}^{\frac{-\Gamma}{l}} - \mathrm{e}^{-\Gamma}$ and number of grasshoppers is represented by N.

Nymph grasshopper's position should alight on ground; hence their position doesn't go below threshold. However, the above-mentioned equation can't be used in simulation of swarm and optimizations in algorithm, reason is that it restrains procedure(algorithm) from exploration and exploitation of the search region (nearby solution). Free space for swarms only can be depicted using this model.

Equation (7) represents the grasshopper interaction in swarm.

The grasshopper rapidly reaches comfort zone and swarm don't merge to a particular point. This causes the optimization problem and the referenced mathematical model becomes unusable(directly). This optimization problem gets solved by modifying the equation a bit. Proposed equation modification is:

$$X_i^d = c\left(\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{ub_d - lb_d}{2} s\left(\left|x_j^d - x_i^d\right|\right)\frac{xj - xi}{d_{ij}}\right) + \widehat{T}_d \tag{8}$$

here $\mathrm{ub_d}$ represents $\mathrm{D_{th}}$ dimension's upper bound, $\mathrm{lb_d}$ represents $\mathrm{D_{th}}$ dimension's lower bound.

$S(\Gamma) = fe^{\frac{-\Gamma}{l}} - \mathrm{e}^{-\Gamma}$, here $\mathrm{D_{th}}$ dimension value in the target (it's the optimized solution arrived as yet) is represented by $\widehat{T}_d$. C refers to reducing coefficient to decrease the attraction, repulsion region and comfort zone. S is similar to S component in Eq. (1).

One important point to be noted is that gravity (no component G) is excluded and underlying assumption is that air direction (i.e. A component) is towards a target ($\underline{T}_d$). Equation (8) depicts next position of grasshopper. It's based on current position of grasshopper, target's position and relative position of all other grass hoppers. First component here is the current grasshopper location relative to all grasshoppers. Current position of all grasshoppers helps in defining the search agent's location around the target. It's the basic difference between Particle Swarm Optimization (PSO) & Grasshopper Optimization algorithm. Most considered swarm intelligent technique is PSO, it involves intelligent collective behaviour of bird's flocks and fish schools. GOA algorithm uses single vector i.e. position where as PSO considers two vectors viz. velocity and position for each particle. In PSO particles don't contribute in revising particle's position. In stark difference, every search agent's next position is determined by all other search agents in GOA.

The adaptive parameters appear two times in the Eq. (8) because of the below reasons:

- Inertial weight (w), which is referenced in PSO, is represented by first C. It decreases the grasshopper movement around the target. This parameter is used to balance Swarm's exploration and exploitation nearby the target.
- The C at second position is used to reduce the attraction, repulsion region and comfort zone. Considering element $c\frac{ub_d - lb_d}{2}s\left(\left|x_j^d - x_i^d\right|\right)$ in the Eq. (8), $c\frac{ub_d - lb_d}{2}$ proportionately reduces the exploration and exploitation region of the grasshoppers. Element $s\left(\left|x_j^d - x_i^d\right|\right)$ indicates whether grasshopper gets repelled from (explore) or attracted to (exploit).

From Eq. (8), the second c (i.e. the inner c) reduces the repulsive/attractive forces between grasshoppers, that is directly proportional to iterations count. However, the first c in Eq. (8) i.e. the outer c, decreases the search coverage near the target as the iteration expands. To summarize, first term in Eq. (8) (summation actually) takes into consideration the location of other grasshoppers and then applies the natural grasshopper interaction.

The other term $\overline{T}_d$ simulates grasshopper's nature of moving in direction of food source. Parameter c represents the slowing down of grasshoppers moving in direction of food. For providing more random behaviour, a random value is multiplied to both terms of equation. Each term can be multiplicated by random values to give random behaviour in either grasshoppers interaction or their food source behaviour.

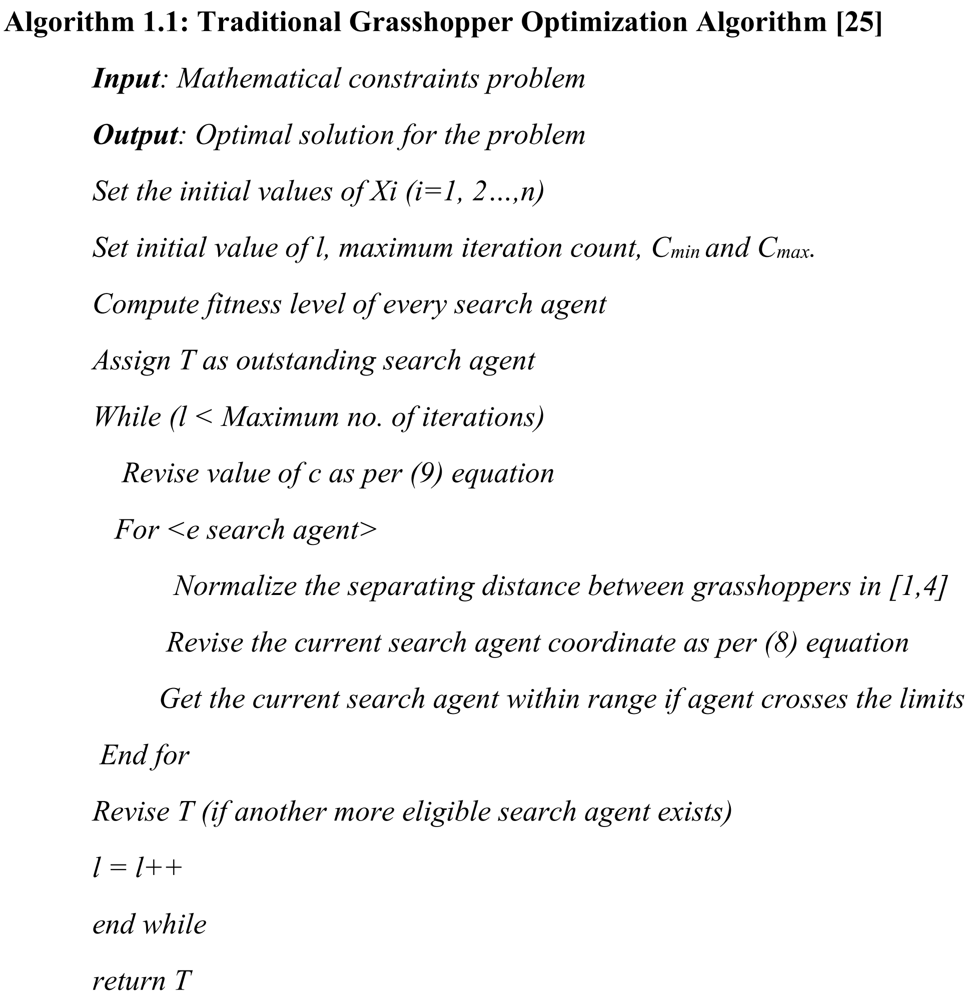
Search space is explored & exploited by this proposed mathematical formulation. A mechanism is required for tuning the level of search agent's exploration to exploitation. In natural form, grasshopper initially is in larvae phase (they have no wings) and hence their initial movement is local for food search. Though during adulthood, there exploration region is large. In stochastic optimization algorithm, this exploration comes first as there need to determine the search space's promising region. Once this promising region is explored, local search is done by exploitation for accurately finding the global optimum.

For getting balance in exploration and exploitation, there exists relation between iterations and number of iterations. Hence the value of c has to be reduced.

Exploitation region increases with the iterations increase. Comfort zone shrinks by c in direct proportion to iterative counter and gets determined as below:

$$C = C_{max} - l\frac{C_{max} - C_{min}}{L} \tag{9}$$

Here $\mathrm{C_{max}}$ represents C's maximum value. $\mathrm{C_{min}}$ represents C's minimum value. l refers to current iteration. Maximum iteration count is L. Here in research calculation, $\mathrm{C_{max}}$ is taken as 1 and and $\mathrm{C_{min}}$ is taken as 0.0001.



Algorithm 1.1: Traditional Grasshopper Optimization Algorithm [25]

```
Input: Mathematical constraints problem
Output: Optimal solution for the problem
Set the initial values of Xi (i=1, 2…,n)
Set initial value of l, maximum iteration count, $C_{min}$ and $C_{max}$.
Compute fitness level of every search agent
Assign T as outstanding search agent
While (l < Maximum no. of iterations)
    Revise value of c as per (9) equation
    For <e search agent>
        Normalize the separating distance between grasshoppers in [1,4]
        Revise the current search agent coordinate as per (8) equation
        Get the current search agent within range if agent crosses the limits
    End for
Revise T (if another more eligible search agent exists)
l = l++
end while
return T
```

2.2 Feature selection

In feature extraction a small set of optimized features is chosen depending on their weight and importance in predicting the end result. Remaining features are neglected. Feature selection is done ensuring the same system performance and as a result giving improved accuracy. Feature set used in ML task is very large number therefore different types of methods are implemented for solving the issue of redundant and irrelevant data feature. Genetic algorithms are many a times improved by using feature selection for reduction in computation costs and better accuracy.

3 Methodology

In this section, the traditional GOA algorithm is enhanced by applying fitness function and sigmoid function resulting in MGOA. MGOA algorithm steps are explained in detail. Secondly, the modified GOA algorithm's flowchart is presented.

3.1 Modified grasshopper optimization algorithm (MGOA)

Presented MGOA algorithm's implementation has been done using Python programming language for feature selection algorithm-based filter method. Below sections detail out the implemented functions.

3.1.1 Init function

It initializes Grasshopper's position in space. Each grasshoppers position is set to 1.

3.1.2 Fitness function

The fitness function involves calculation of each agent's fitness using the following equation and then returns the fitness value of each grasshopper:

$$fitness\_goa = imp + bf \times (1 - (selected\_features/total\_features)) \tag{10}$$

In this equation, "imp" is sum of each selected feature's importance. Feature importance is a characteristic of Random Forest Classifier from sklearn library. Feature importance produces value in the range of 0 to 1 for every feature as per their importance in target prediction. Sum of all feature's importance is always equating to unity.

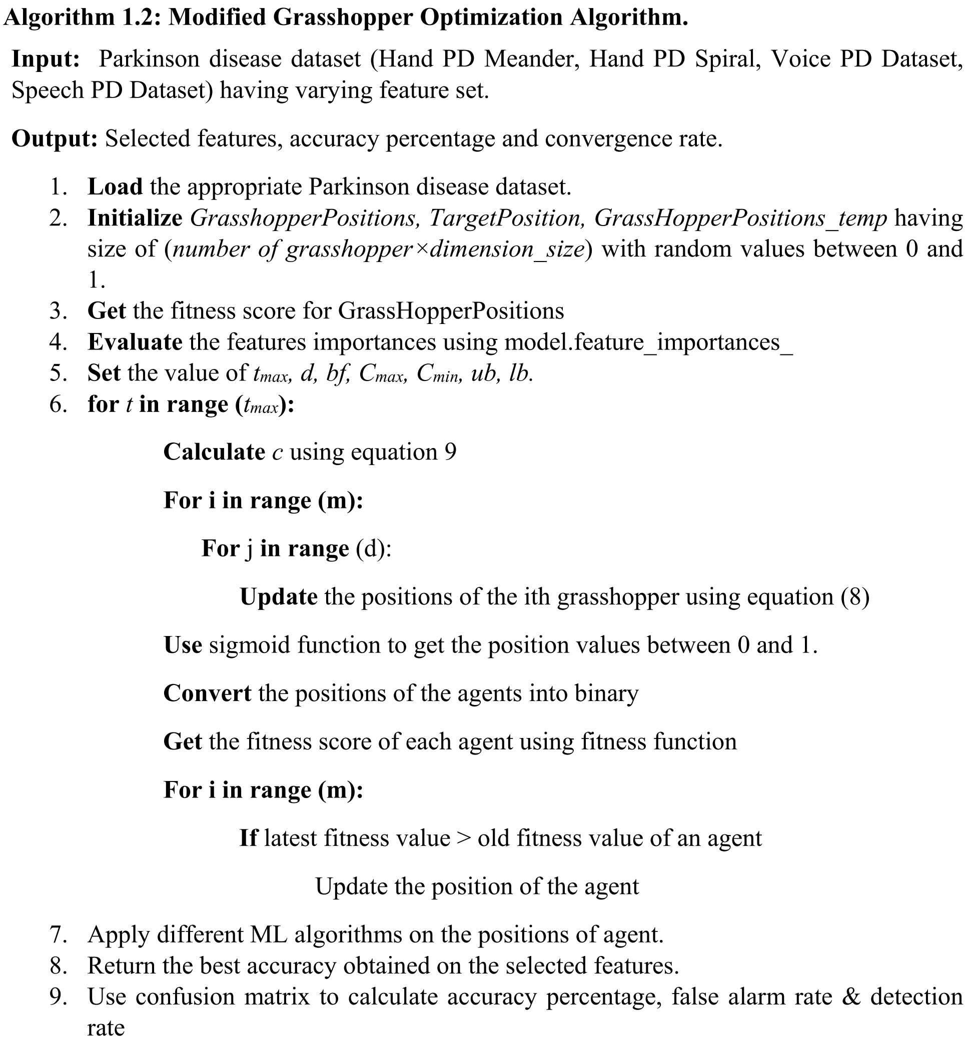
bf represents the balancing factor, which generates-implemented on all four Parkinson Disease a balance between the feature importance, imp and selected features count.

selected_feature is the count of shortlisted features.

total_feature is the count of all feature in dataset.

A sigmoid function is an activation function, also known as squashing function. It limits the output between 0 and 1and is used in the prediction of probabilities.

Ultimate goal is to minimize the value of this fitness function for finding the improved solution for detection of the Parkinson's disease. The presented Modified Grasshopper Optimization Algorithm (MGOA) processes feature set as input and provides output as the reduced feature set (subset) that upgrades the model performance. MGOA pseudo code is enlisted in the below section titled Algorithm 1.2.



Algorithm 1.2: Modified Grasshopper Optimization Algorithm.

Input: Parkinson disease dataset (Hand PD Meander, Hand PD Spiral, Voice PD Dataset, Speech PD Dataset) having varying feature set.

Output: Selected features, accuracy percentage and convergence rate.

```
1. Load the appropriate Parkinson disease dataset.
2. Initialize GrasshopperPositions, TargetPosition, GrassHopperPositions_temp having
   size of (number of grasshopper×dimension_size) with random values between 0 and
   1.
3. Get the fitness score for GrassHopperPositions
4. Evaluate the features importances using model.feature_importances_
5. Set the value of $t_{max}$, $d$, $bf$, $C_{max}$, $C_{min}$, $ub$, $lb$.
6. for $t$ in range ($t_{max}$):
       Calculate $c$ using equation 9
       For i in range (m):
           For j in range (d):
               Update the positions of the ith grasshopper using equation (8)
       Use sigmoid function to get the position values between 0 and 1.
       Convert the positions of the agents into binary
       Get the fitness score of each agent using fitness function
       For i in range (m):
               If latest fitness value > old fitness value of an agent
                       Update the position of the agent
7. Apply different ML algorithms on the positions of agent.
8. Return the best accuracy obtained on the selected features.
9. Use confusion matrix to calculate accuracy percentage, false alarm rate & detection
   rate
```

Table 1 Input parameters

Parameters	Values	Description
T_{max}	50	Total iterations count
Dim	Feature count in dataset	Count of all features(dimensions)
lb	0	Lower boundary limit
ub	10	Upper boundary limit
cMax	1	
cMin	0.00001	
bf	1.0	Balancing factor (balances the weight between feature importance and selected features)

The present algorithm used for detection of Parkinson's ailment have been optimized to improve the performance. The updated algorithm changes are enumerated as follows:

- The modified grasshopper optimization algorithm is designed for multiple feature shortlisting which extends the original GOA (it's implemented by using mathematical function).
- The fitness function involves calculation of each agent's fitness using the Eq. (10) and then returns the fitness value of each grasshopper.
- The co-ordinates of the grasshoppers are updated based on the Eq. 8.
- For converting the agent's position value between 0 and 1, the sigmoid function is used. The equation is given by

$$sigmoid = \frac{1}{(1 + \exp((10 * (X - 0.5))))} \tag{11}$$

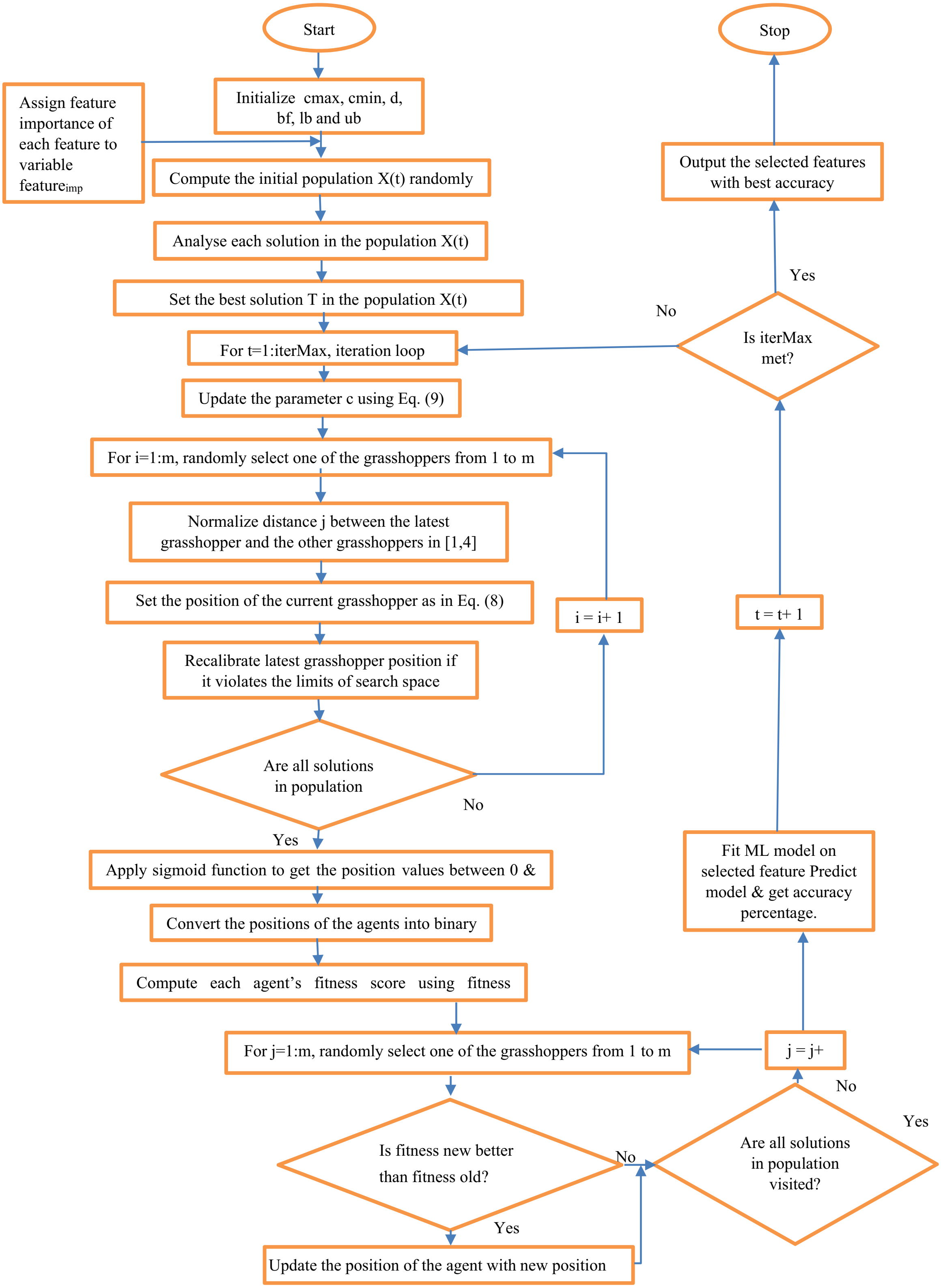
- Convert the positions to binary values between 0 and 1 so they can be used for feature selection. Unselected feature is assigned 0 value and selected feature is assigned 1.

Table 2 Tuning parameters of Machine Learning models

Model	Parameters
Decision trees	max_depth having value 30 min_sample_split with value 20
Random forest	n_estimators with value 500
k-NN	n_neighbors having value 3

- The new positions obtained is passed through fitness function to get fitness score.
- The Target position is updated with the values of new position based on the condition if new calculated fitness is better than old fitness value.
- The positions after the t_{max} iterations is utilized for the training and validating various ML algorithms: Decision Tree, Random Forest, k-NN are applied on the chosen positions. False alarm rate, accuracy percentage and detection rate are obtained by using the confusion matrix. Improved prediction is the desired result.

Flow Chart for **Modified Grasshopper Optimization Algorithm** is given below


Start
Initialize cmax, cmin, d, bf, lb and ub
Assign feature importance of each feature to variable feature$_{imp}$
Compute the initial population X(t) randomly
Analyse each solution in the population X(t)
Set the best solution T in the population X(t)
For t=1:iterMax, iteration loop
Update the parameter c using Eq. (9)
For i=1:m, randomly select one of the grasshoppers from 1 to m
Normalize distance j between the latest grasshopper and the other grasshoppers in [1,4]
Set the position of the current grasshopper as in Eq. (8)
Recalibrate latest grasshopper position if it violates the limits of search space
Are all solutions in population
No
i = i+ 1
Yes
Apply sigmoid function to get the position values between 0 &
Convert the positions of the agents into binary
Compute each agent's fitness score using fitness
For j=1:m, randomly select one of the grasshoppers from 1 to m
Is fitness new better than fitness old?
Yes
Update the position of the agent with new position
No
Are all solutions in population visited?
No
j = j+
Yes
Fit ML model on selected feature Predict model & get accuracy percentage.
t = t+ 1
Is iterMax met?
No
Yes
Output the selected features with best accuracy
Stop

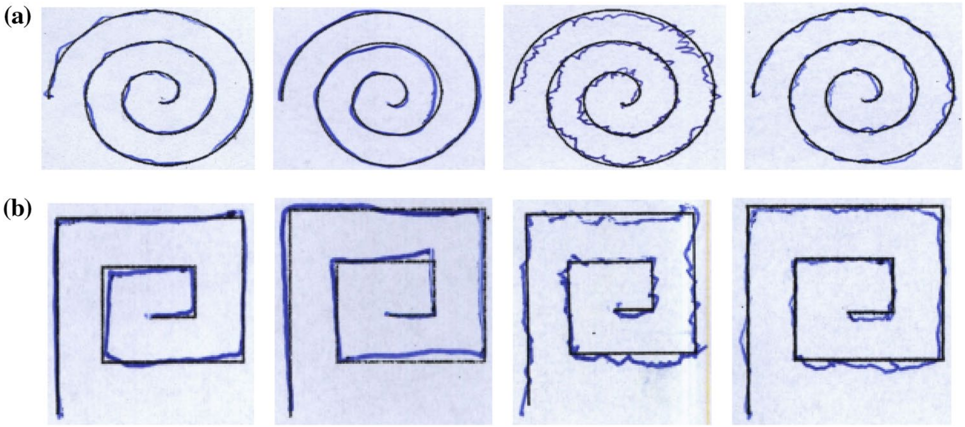


Fig. 2 **a** Hand PD Spiral Dataset, **b** Hand PD Meander Dataset

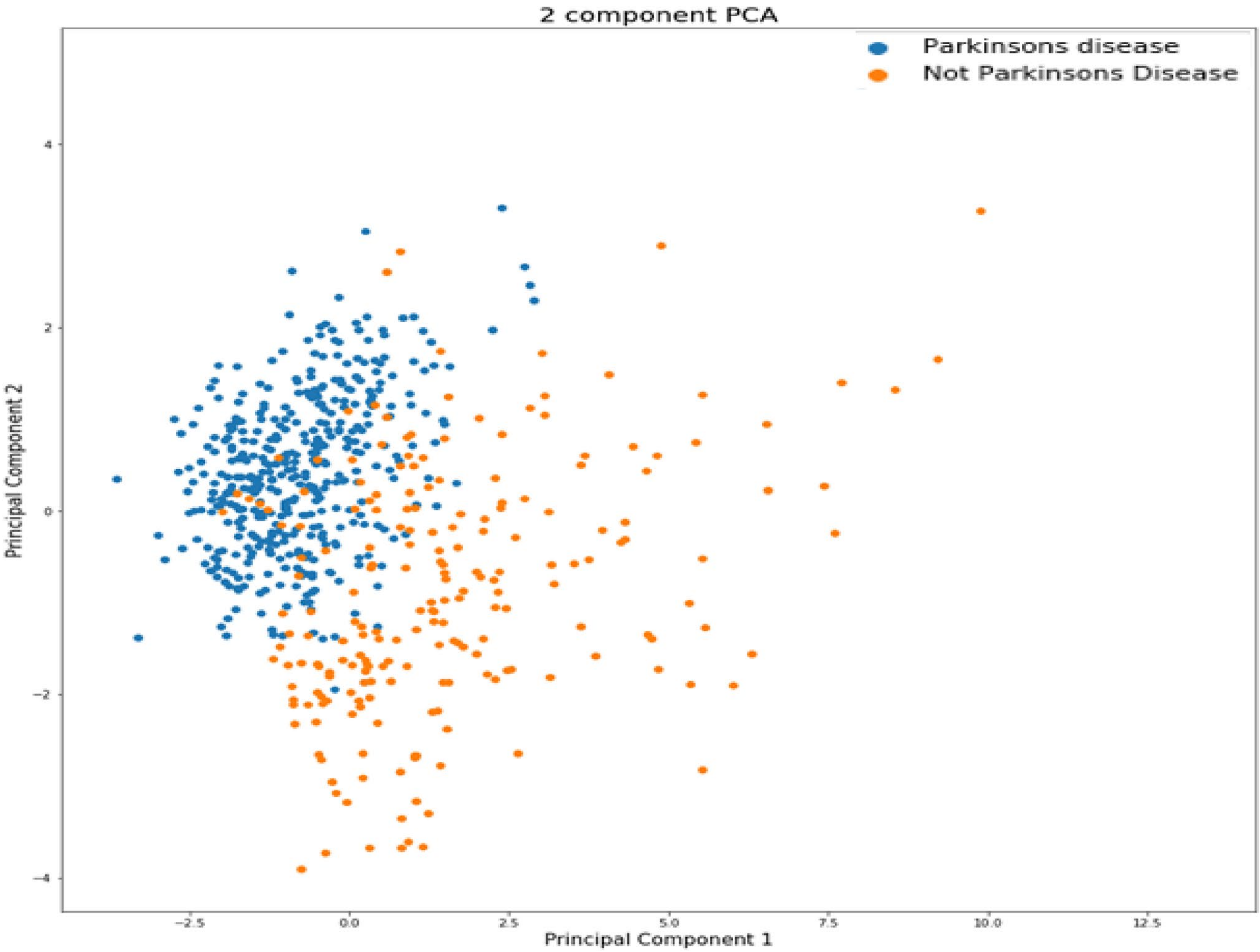


Fig. 3 Hand PD spiral dataset—Scatter plot

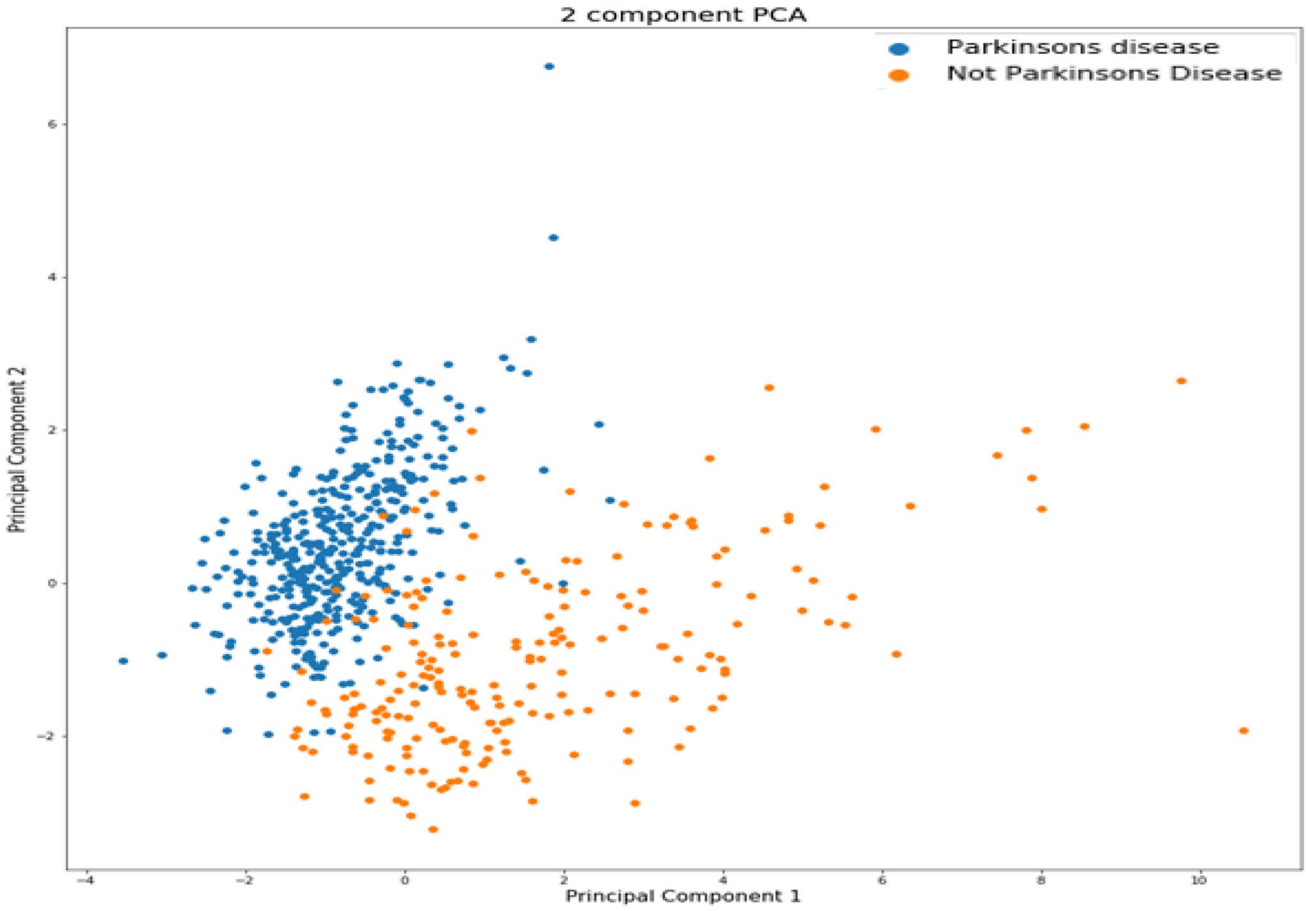


Fig. 4 Hand PD meander dataset—Scatter plot

Table 3 Features selected From Hand PD (a) meander, (b) spiral by MGOA

S. no	Feature selected (Name)
a	
1	GENDER
2	RIGH/LEFT-HANDED
3	AGE
4	MAX_BETWEEN_ET_HT
5	MIN_BETWEEN_ET_HT
6	STD_DEVIATION_ET_HT
7	MRT
8	STD_HT
b	
1	AGE
2	MRT
3	MIN_HT
4	STD_HT
5	CHANGES_FROM_NEGATIVE_TO_POSITIVE_BETWEEN_ET_HT

4 Implementation of the proposed algorithm

This section details the required setup for experimentation, the inputs require i.e. parameters, ML algorithms, the optimized parameters and all the data sets.

4.1 Experimental setup

System configuration of Intel® Core™ i5-7200U and CPU of 2.50 GHz × 4 under Operating system Ubuntu 18.04(version) was used for testing the proposed algorithm. The algorithm has been implemented using Python 3.6.3(version): Anaconda, Inc.

4.2 Input parameters

Table 1 has description of input parameters.

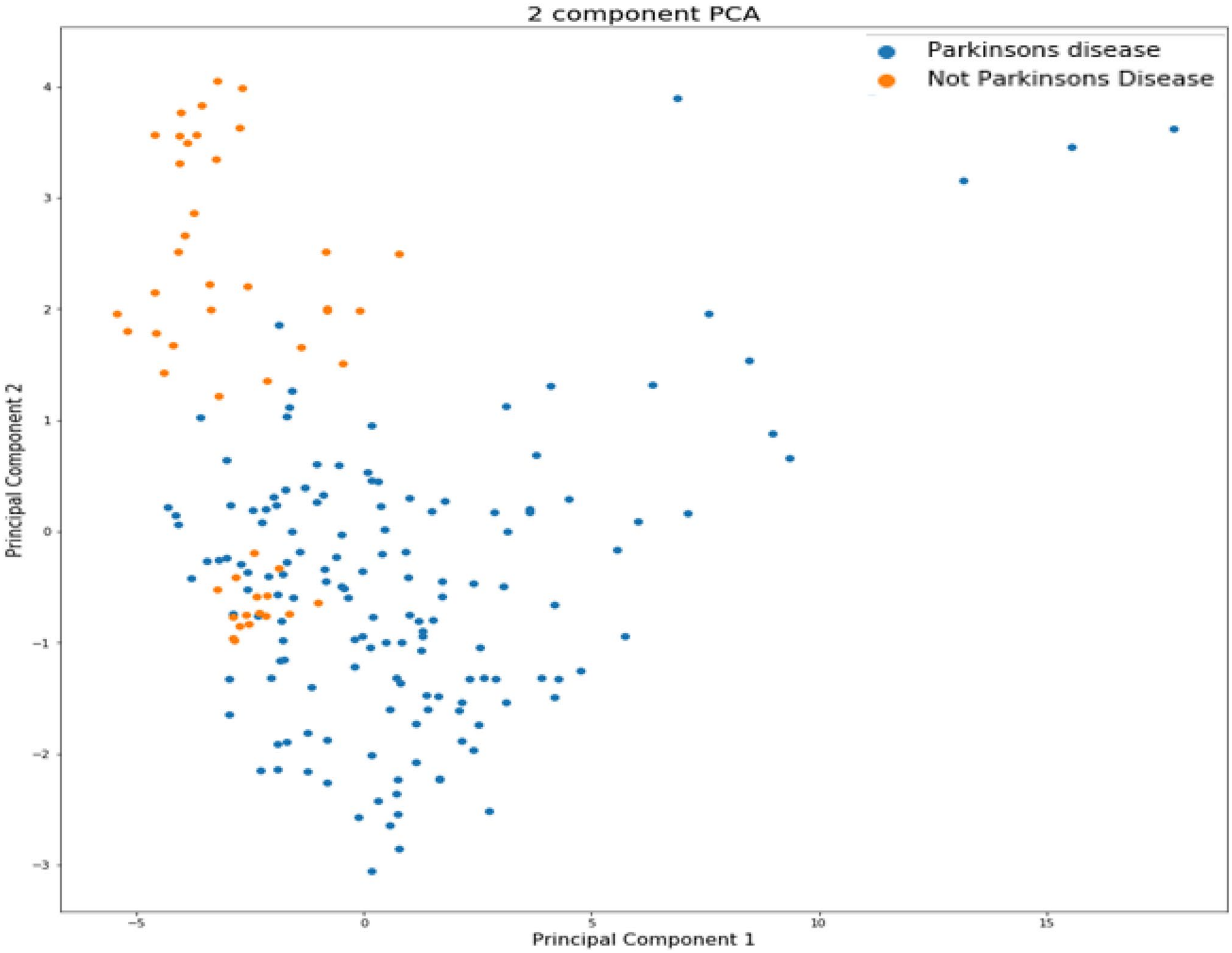


Fig. 5 Speech PD dataset Scatter plot

4.3 Machine learning models

There are three popular Machine Learning Models Viz. Random forest, k-Nearest Neighbour(k-NN) and decision tree. These models are utilized for the visualization of target. For our experiment we have data is split into 20–80 ratio for testing and training and then shambled so that overfitting does not happen. Problem of classification and regression is solved well by Random forest which forms tree forest. In problems related to classification, supervised learning algorithm k-NN is used. Here a positive odd integer K classifies object to the most frequently used class in the neighbourhood. We have taken value of k as 3 in our research. This parameter k is tuned for accuracy. Another machine learning algorithm that's used for solving problems related to regression and classification is Decision tree. In Decision tree method trees are used for dividing the population in sub population [31].

Table 2 below lists tuning parameters of all these three machine learning algorithms.

4.4 Datasets

4.4.1 Parkinson datasets

Different Parkinson Disease datasets are compiled from handwriting tests to recordings of sound (audio), on these datasets GOA algorithm is applied. Details of these dataset are given as follows:

4.4.1.1 Hand PD dataset Handwritten tests of patients and healthy individuals comprise of this dataset. The participants give the required information by filling a form for the purpose of research and by drawing spirals and meanders. The total participants are 158 out of which 53 are healthy and 105 are patients. Average of healthy par-

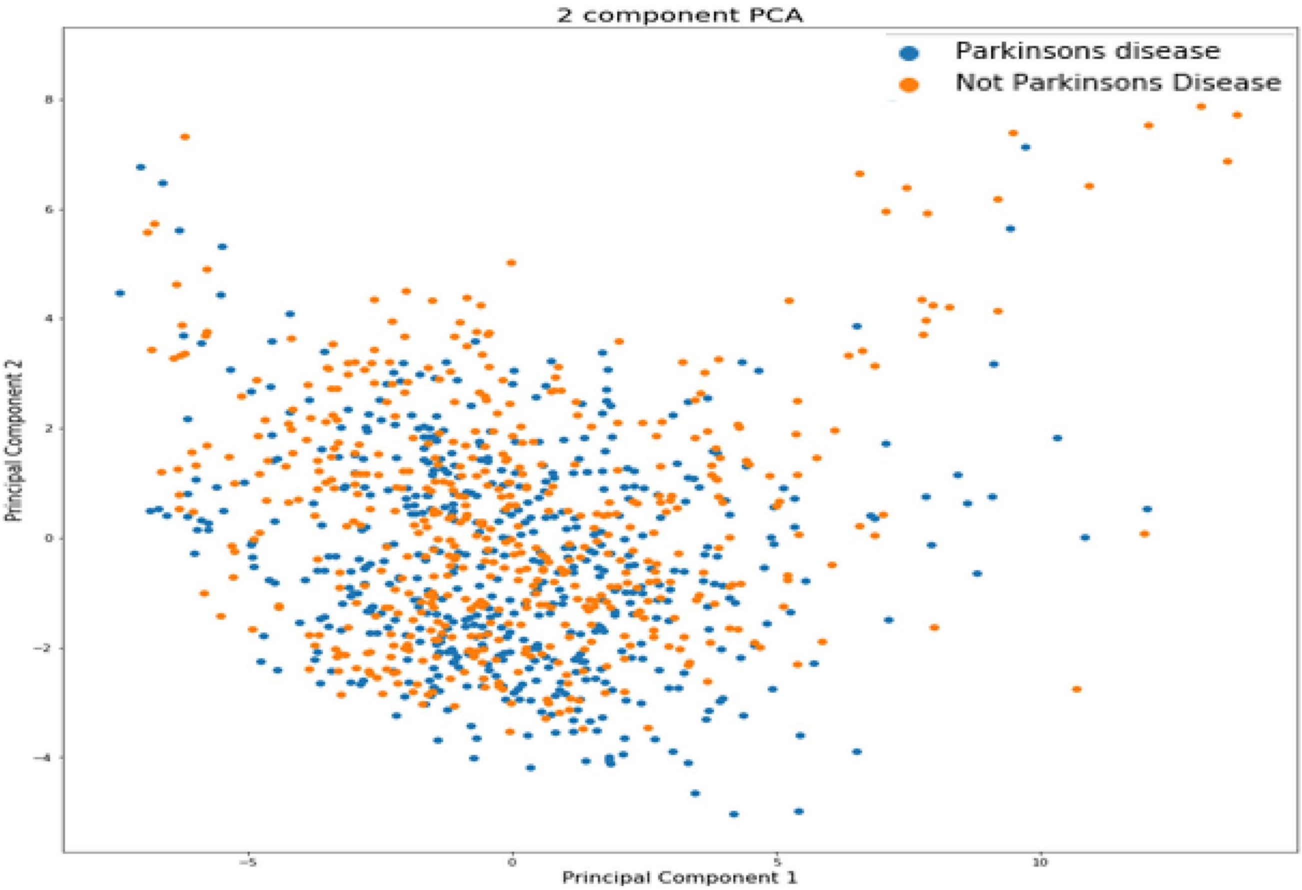


Fig. 6 Voice PD data—Scatter Plot

ticipant's age is 44.22 ± 16.53 years and average patients age 58.75 ± 7.51 years.

As shown in Fig. 2a, b, each participant draws 4 spirals and 4 meanders. In total there are 632 instances and 13 features. For studying the Parkinson disease progress over time in Brazil and for analysing different stages during the progress [2], a research was carried out at BMS (Botucatu Medical School), SP (São Paulo) State University. And in this research handwritten tests were done.

Figures 3 and 4 have scatter plot for meander dataset and hand PD spiral.

Dataset characteristics are:

- Problem Type: Classification
- Datasets Characteristic: Multivariable
- Attributes/features Characteristics: real
- Count of Instances: 632
- Count of features: 13
- Feature targeted: CLASS_TYPE
- Values missing: N/A

On application of MGOA on Hand PD dataset, Eight and five features got selected out of 13 which have been listed in Table 3 a, b.

5 Speech PD dataset

Biomedical voice (audio) measurement samples were taken from the recordings of voice (audio) of 31 participants, in which actually 23 participants had PD (Parkinson's disease). In Total 195 voice (audio) recordings were generated. Dataset was created in such a way that every column gets a specific voice measurement for all the 195 rows (different recordings). This data was primarily used for selecting features which had most contribution in PD, hence differentiating PD and healthy participants. Status column has 0 corresponding for healthy participants and 1 for PD participants. In data set, individual ages vary between 46 and 85 years. Recording of speech signals was done by Sir Max Little (from University of Oxford), in association with National Centre for Voice & Speechka at

Table 4 Results found by MGOA when applied on (a): HandPD spiral dataset, (b) HandPD Meander Dataset, (c) Speech PD Dataset, (d) voice PD Dataset

Classifiers	Accuracy	Detection Rate	False Alarm Rate
a			
MGOA- KNN	75.59	85.26	53.12
MGOA using Random Forest	**92.91**	**97.89**	**21.87**
MGOA using Decision Tree	88.97	94.73	28.12
b			
MGOA- KNN	74.8	85.8	47.62
MGOA- Random Forest	**93.7**	**100**	19.05
MGOA-Decision Tree	88.98	91.76	**16.67**
c			
MGOA-KNN	89.74	96.67	30
MGOA-Random Forest	**94.87**	**100**	**22.22**
MGOA-Decision Tree	84.61	90	30
d			
MGOA-KNN	91.82	83.51	0.91
MGOA-Random Forest	**100**	**100**	**0**
MGOA-Decision Tree	100	100	0

Bold values indicate best result compared to others

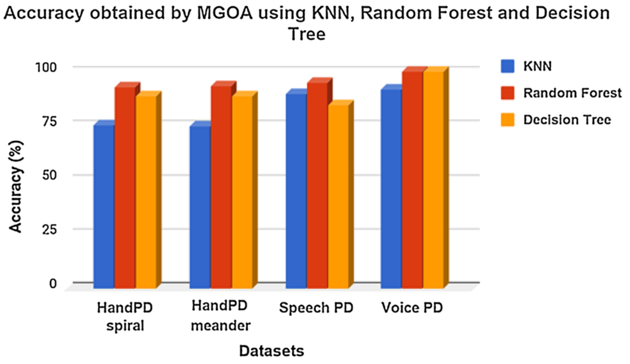


Fig. 7 Accuracy obtained by MGOA with Random Forest, KNN & Decision tree

Denver, Colorado [3] contributed in generating this dataset. Figure 5 depicts the speech PD dataset's scatter plot.

Dataset characteristics are enumerated below:

- Problem Type: Classification
- Dataset characteristics: multivariable
- Attributes/features characteristics: real
- Count of Instances: 195
- Count of Features: 23
- Target feature: Status
- Missing values: N/A

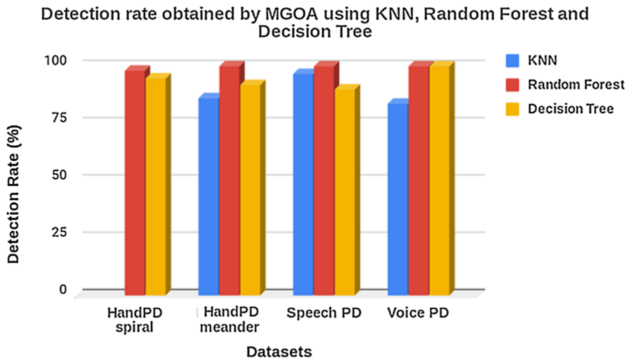


Fig. 8 Detection rate of MGOA with Random Forest, KNN & Decision tree

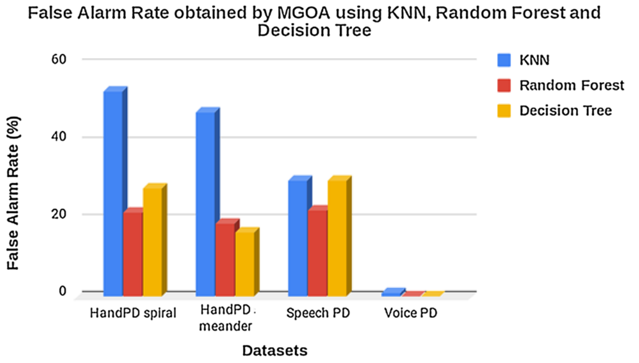


Fig. 9 False alarm rate of MGOA with Random Forest, KNN & Decision tree

6 Voice PD dataset

Voice of 40 participants who appeared at the Neurology Department in Cerrahpasa Faculty of Medicine, Istanbul University [1] constitutes this data set. Out of these 40 participants 20 are patients and 20 are healthy. Out of 20 patients, 14 are male and 6 are female. Whereas out of 20 healthy participants, 10 are male and 10 are female.

Participants generated 26 sound recording samples (constituting numbers, words, vowels, sentences). There are two distinct files viz. testing and training in this data set. There are many voice(audio) recordings are there in the training set. 28 Parkinson disease patients were given instruction to pronounce say vowels 'a' and 'o' thrice (28 × 6 = 168), there by generating a sample of 168 samples making up the test set. This test set was used for validating the model that was trained by using training set. Figure 6 depicts the scatter plot for voice PD dataset. The 26 feature sets are described in [1].

Datasets main characteristics are:

- Problem type: Classification

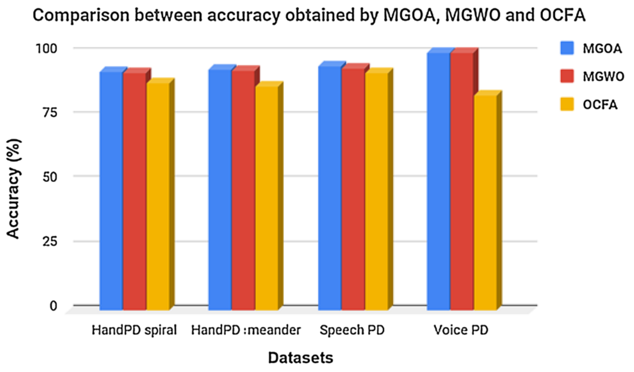


Fig. 10 Comparison between Accuracy percentage of MGOA, MGWO and OCFA

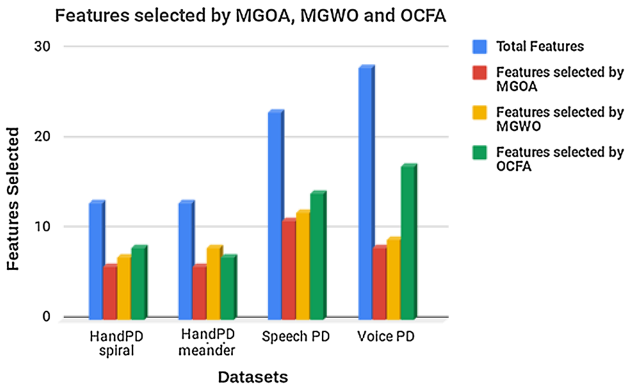


Fig. 11 Comparison between selected features by MGOA, MGWO and OCFA

- Datasets characteristic: multivariable
- Attributes/features characteristics: real
- Count of Instances: 1040
- Count of Features: 26
- Missing values: N/A

7 Results and discussions

The presented algorithm is applied on various Parkinson disease datasets (Hand PD meander, Hand PD spiral, Voice PD and Speech PD) and the obtained results are discussed here in this section. Also, the novel and new algorithm MGOA is contrasted with Optimized Cuttlefish Algorithm (OCFA) [32] and Modified Grey Wolf Optimizer (MGWO) [33]. This section contains accuracy attained by MGOA (using classifier KNN, Random Forest, Decision tree) when applied to four datasets then MGOA's improved accuracy has been contrasted with the accuracy of OCFA and MGWO for every data set. The features shortlisted by MGOA, OCFA and MGOA are also compared. In the end the convergence rate of the proposed algorithm for all the datasets is shown. Classifier performance is considered good when AR, DR values are on higher side where as FAR values are on lower side. On applying the MGWO classifiers to all datasets, performance metrics utilized to appraise the model are:

Accuracy (AR) is how often the classifier correctly recognizes diseased as well as normal individuals.

$$AR = \frac{No.\ of\ Correctly\ recognized\ cases}{Total\ number\ of\ cases\ in\ the\ testset} * 100$$

Detection rate (DR) is how many times the classifier model identifies a Parkinson patient correctly.

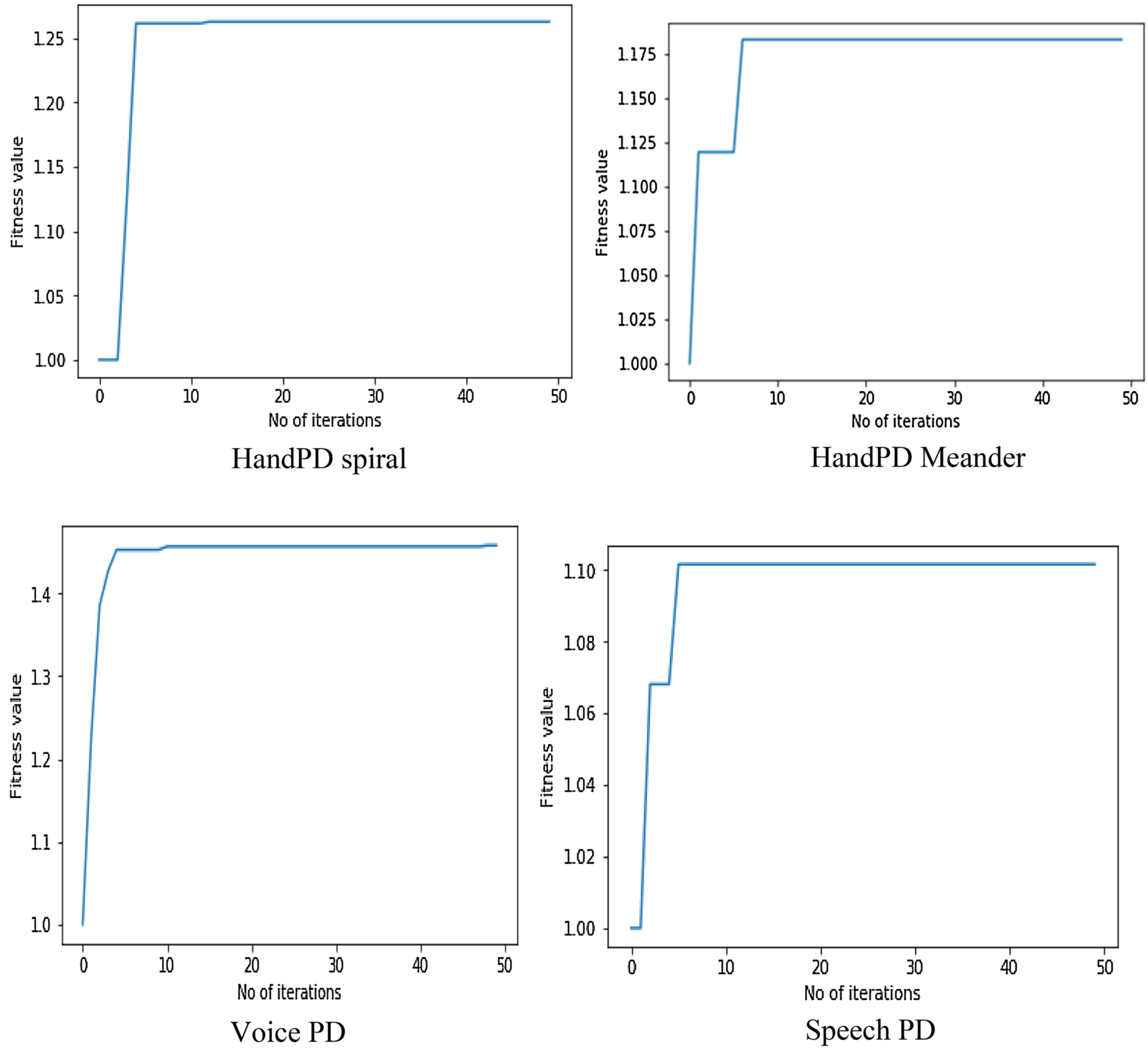
$$DR = \frac{No.\ of\ correctly\ recognised\ patient}{Total\ no.\ of\ parkinson\ patient\ in\ the\ test\ set} * 100$$

False alarm rate (FAR) is number of times falsely identified as a Parkinson patient out of all normal individual in test set.

$$FAR = \frac{No.\ of\ normal\ individuals\ falsely\ recognized\ as\ parkinson\ patients}{Total\ no.\ of\ normal\ individual\ in\ the\ test\ set} * 100$$

In this algorithm three different popular ML models have been utilized: Decision Tree, Random Forest and KNN. AR, DR and FAR values for the new model when implemented on all four Parkinson Disease datasets (Hand PD meander, Hand PD spiral, Voice PD and Speech PD) are shown in Table 4a, b, c, d.

Table 4a shows that Accuracy and Detection rate is highest and False Alarm Rate is least for classifier MGOA using Random Forest.



HandPD spiral

HandPD Meander

Voice PD

Speech PD

Fig. 12 Convergence rate for all datasets

Table 5 MGOA-Random Forest's Average Accuracy percentage, Detection Rate & False Alarm Rate

Classifiers	Accuracy (Avg.)	Detection rate (Avg.)	False alarm rate (Avg.)
MGOA-Random Forest	**95.37**	**99.47**	**15.78**

From the above results, it's evident that Accuracy and Detection rate is highest for classifier MGOA using Random Forest.

Study results show that Accuracy and Detection rate is highest and False Alarm Rate is least for classifier MGOA using Random Forest. Detection rate is 100% in this case.

Results show that Accuracy percentage and Detection rate are 100% and False Alarm rate is 0 for classifier MGOA-Random Forest in case of Voice PD Dataset.

From the Table 4 no. 2 accuracy percentage, detection rate and false alarm rate for these are depicted in Figs. 7, 8 and 9 respectively. It can be concluded from the Figs. 7, 8 and 9 that MGOA using Random Forest is giving the improved results

8 Comparison between MGOA, MGWO & OCFA

Accuracy results found by shortlisted features of MGOA using Random forest has been contrasted with the accuracy of MGWO [33] and OCFA [32]. The comparison between the three algorithms is shown below in Figs. 10 and 11. From the below figure, it's evident that Out of the three algorithms MGOA is more accurate and has smallest count of shortlisted features. Below Fig. 10 shows accuracy comparison & Fig. 11 shows the comparison of shortlisted features between the algorithms.

Statistical significance of tabulated results in Table 4a-d and the associated graphs in Figs. 7, 8, 9, 10 and 11 show that presented MGOA algorithm clearly has better accuracy percentage, detection rate and has least false alarm rate. Their computed average has been tabulated in Sect. 5 below.

Datasets convergence rate shows the algorithm convergence towards the global optima [34]. As shown in Fig. 12.

Prime reason of achieving low runtime by the proposed algorithm MGOA is quick convergence ability, even when the search space is unknown. In addition, MGOA enabled fast convergence by specifying the number of main iterations.

9 Conclusion and future work

This research work recommends a modified form of bio-inspired Grasshopper Optimization Algorithm for selection of features. A reduced feature set is found by using Modified GOA. The proposed algorithm mimics the tendency of the grasshopper swarms and their social interaction and has been used for solving real world problems. The presented algorithm is detailed in Sect. 3. Proposed MGOA has been applied on various datasets for Parkinson's ailment, whose spatial distribution is depicted in Figs. 3, 4, 5, and 6. The comparative accuracy of the proposed algorithm with computational time is found without compromising on model's performance measure.

Presented algorithm shortlists a smaller feature set in comparison to OCFA [32] & MGWO [33]. As listed in Table 5, it has better accuracy of 95.37% with detection rate of 99.47% and false alarm rate of 15.78%. Performance of MGOA with Random Forest is better than other ML classifiers i.e. decision tree, and K-NN. Presented algorithm can be put to use for early detection of Parkinson's disease and also for feature shortlisting.

In future, advanced research can be carried out by creating new algorithms for combining the models of voice PD, Speech PD and Hand PD datasets so that early diagnosis can be done with improved accuracy. Other than Voice, Speech and Hand methods, more methods can be tested with the presented algorithm for validation of results. MGOA has been implemented for Parkinson disease optimization problem, in the same way it can be used for other optimization problems for getting improved results. This newly designed algorithm application on Parkinson's Image dataset could be done.

Compliance with ethical standards

Conflict of interest The author(s) declare that they have no competing interests.

References

1. Erdogdu Sakar B, Isenkul M, Sakar CO, Sertbas A, Gurgen F, Delil S, Apaydin H, Kursun O (2013) Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inf 17(4):828–834
2. Pereira CR, Pereira DR, Silva FA, Masieiro JP, Weber SAT, Hook C, Papa JP (2016) A new computer vision-based approach to aid the diagnosis of Parkinson's disease. Comput Methods Programs Biomed 136:79–88
3. Little MA, McSharry PE, Hunter EJ, Ramig LO (2008) Suitability of dysphonia measurements for telemonitoring of Parkinson 's disease. IEEE Trans Biomed Eng 56(4):1015–1022
4. Pereira CR, Pereira DR, Weber SAT, Hook C, de Albuquerque VHC, Papa JP (2018) A survey on computer-assisted Parkinson's Disease diagnosis. Artificial Intelligence in Medicine.
5. Lakshman PSK, Shankar K, Khanna A, Gupta D, Rodrigues JJPC, Pinheiro PR, Albuquerque VHCD (2018) Effective feature to classify big data using social internet of things. IEEE Access 6:24196–24204
6. Keswania B, Ambarish Mohapatra A, Mohanty A, Khanna A, Rodrigues J, Gupta D, de Albuquerque VHC (2018) Adapting weather conditions based IoT enabled smart irrigation technique in precision agriculture mechanisms. Neural Comput Appl 20:1–16
7. Rodrigues R, Rodrigues JJPC, Cruz M, Khanna A, Gupta D (2018) An IoT-based automated shower system for smart homes. In: International conference on advances in computing, communications, and informatics (ICACCI'18), pp 254–258
8. Nayyar A, Garg S, Gupta D, Khanna A (2018) Evolutionary computation- theory and algorithms. In: Advances in swarm intelligence for optimizing problems in computer science. CRC Press, Boca Raton, pp 1–23, chapter 1
9. Nakamura RYM, Pereira L, Costa K, Yang XS. BBA: a binary bat algorithm for feature selection. Brazil: Department of Computing Sao Paulo State University B̂ auru; 2012
10. Gupta D, Ahlawat AK (2017) Usability feature selection via MBBAT: a novel approach. J Comput Sci 23:195–203
11. Prior H, Schwarz A, Güntürkün O (2008) Mirror-induced behavior in the magpie (pica pica): evidence of self-recognition. PLoS Biol 6(8):e202
12. Xin-She Y (2009) Firefly algorithms for multimodal optimization. Stochastic Algorithms: Foundations and Applications, SAGA, Lecture Notes in Computer Sciences 5792:169–178
13. Xin-She Y (2010) "A new metaheuristic bat-inspired algorithm", NICSO. SCI, Springer, Heidelberg 284:65–74

14. Gonzalez-Pardo A, Jung JJ, Camacho D (2017) ACO-based clustering for ego network analysis. Future Gener Comput Syst 66:160–170
15. Pham DT, Ghanbarzadeh A, Koc E, Otri S, Rahim S, Zaidi M (2006) The bees algorithm-a novel tool for complex optimization problems. In: Proceedings of the 2nd virtual international conference on intelligent production machines and systems (IPROMS 2006), pp 454–45
16. Erik C, Mauricio G, Daniel Z, Marco PC, Guillermo G (2012) An algorithm for global optimization inspired by collective animal behavior, Hindawi Publishing Corporation Discrete Dynamics in Nature and society Article ID 638275
17. Shankar K, Lakshmanaprabu SK, Khanna A, Tanwar S, Rodrigues JJPC, Ranjan Roy N (2019) Alzheimer detection using Group Grey Wolf Optimization based features with convolutional classifier. Comput Electr Eng 77:230–243
18. Gupta N, Gupta D, Khanna A, Rebouças Filho PP, de Albuquerque VHC (2019) Evolutionary algorithms for automatic lung disease detection. Measurement 140:590–608
19. Gupta D, Ahlawat A, Sagar K (2014) A critical analysis of a hierarchy-based usability model. In: International conference on proc. Int. Conf. Contemp. Comput. Inform. (ICI), 2014. IEEE, pp 255–260
20. Gupta D, Rodrigues JJPC, Sundaram S, Khanna A, Korotaev V, Albuquerque VHC (2018) Usability feature extraction using modified crow search algorithm: a novel approach. Neural Comput Appl 68:412–424
21. Jain R, Gupta D, Khanna A (2018) Usability feature optimization using MWOA. In: International conference on innovative computing and communication (ICICC), vol 56, pp 453, 2019
22. Shankar K, Lakshmanaprabu SK, Gupta D, Maseleno A, de Albuquerque VHC (2018) Optimal features-based multi-kernel SVM approach for thyroid disease classification. J Supercomput 1:1–16
23. Gupta D, Sagar K (2010) Remote file synchronization single-round algorithm. Int J Comput Appl 4(1):32–36
24. Patnaik A, Gupta D (2010) Unique identification system. Int J Comput Appl 7(5):46–51
25. Saremi S, Mirjalili S, Lewis A (2017) Grasshopper optimisation algorithm: theory and application. Adv Eng Software 105:30–47
26. Tumuluru P, Ravi B (2018) Chronological grasshopper optimization algorithm-based gene selection and cancer classification. J Adv Res Dyn Control Syst 10(3):80–94
27. Ibrahim HT, Mazher WJ, Ucan ON, Bayat O (2018) A grasshopper optimizer approach for feature selection and optimizing SVM parameters utilizing real biomedical data sets. Neural Comput Appl, 1–10
28. Simpson SJ, McCaffery AR, Hagele BF (1999) A behavioural analysis of phase change in the desert locust. Biol Rev 74:461–480
29. Rogers SM, Matheson T, Despland E, Dodgson T, Burrows M, Simpson SJ (2003) SimpsonMechanosensory-induced behavioural gregarization in the desert locust Schistocerca gregaria. J Exp Biol 206:3991–4002
30. Topaz CM, Bernoff AJ, Logan S, Toolson W (2008) A model for rolling swarms of locusts. Eur Phys J Spec Top 157:93–109
31. Sehgal S, Agarwal M, Gupta D, Bashambu A (2019) Comparative Study on machine learning and data mining techniques for diseases diagnosis. Int J Manag IT Eng vol 9(1), ISSN: 2249–0558.
32. Gupta D, Julka A, Jain S et al (2018) Optimized cuttlefish algorithm for diagnosis of Parkinson's disease. Cognit Syst Res 52:36–48
33. Sharma P, Sundaram S, Sharma M, Sharma A, Gupta D (2019) Diagnosis of Parkinson's disease using modified grey wolf optimization. Cognit Syst Res 54:100–115
34. Mirjalili S, Mirjalili SM, Hatamlou A (2016) Multi-verse optimizer: a nature-inspired algorithm for global optimization. Neural Comput Appl 27(2):495–513

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.