

A new parallel multi-objective Harris hawk algorithm for predicting the mortality of COVID-19 patients (#82654)

1

First submission

Guidance from your Editor

Please submit by **28 Mar 2023** for the benefit of the authors (and your token reward) .



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Raw data check

Review the raw data.



Image check

Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the [materials page](#).

4 Raw data file(s)

1 Other file(s)



Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).

BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [PeerJ standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [PeerJ policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  All underlying data have been provided; they are robust, statistically sound, & controlled.
-  Conclusions are well stated, linked to original research question & limited to supporting results.



The best reviewers use these techniques

Tip

Example

Support criticisms with evidence from the text or from other sources

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

A new parallel multi-objective Harris hawk algorithm for predicting the mortality of COVID-19 patients

Tansel Dokeroglu Corresp. 1

¹ Cankaya University, Software Engineering Department, Ankara, Turkey

Corresponding Author: Tansel Dokeroglu
Email address: tanseldoker@gmail.com

Harris' Hawks Optimization (HHO) is a novel metaheuristic inspired by the collective hunting behaviours of hawks. This technique employs the patterns of hawks to produce (near)-optimal solutions for challenging classification problems that are enhanced with feature selection. In this study, we propose a new parallel multi-objective HHO algorithm for predicting the mortality of COVID-19 patients according to their symptoms. There are two objectives in this optimization problem; to reduce the number of features and increase the accuracy of the predictions. We conduct comprehensive experiments on a recent real-world COVID-19 dataset from Kaggle. An augmented data version of the COVID-19 dataset is also generated and analyzed to improve the quality of the solutions. Significant improvements are observed compared to state-of-the-art metaheuristic wrapper algorithms. We reported better classification results with feature selection than using the whole set of features of the dataset. 98.15% prediction accuracy is achieved with a 45% reduced number of features during experiments. We were successful in obtaining the new best solutions in the literature for the COVID-19 dataset.

A new Parallel Multi-objective Harris Hawk Algorithm for Predicting the Mortality of COVID-19 Patients

Tansel Dokeroglu¹

¹Department of Software Engineering, Cankaya University, Ankara, Turkey

Corresponding author:

Tansel Dokeroglu¹

Email address: tdokeroglu@cankaya.edu.tr

ABSTRACT

Harris' Hawks Optimization (HHO) is a novel metaheuristic inspired by the collective hunting behaviours of hawks. This technique employs the patterns of hawks to produce (near)-optimal solutions for challenging classification problems that are enhanced with feature selection. In this study, we propose a new parallel multi-objective HHO algorithm for predicting the mortality of COVID-19 patients according to their symptoms. There are two objectives in this optimization problem; to reduce the number of features and increase the accuracy of the predictions. We conduct comprehensive experiments on a recent real-world COVID-19 dataset from Kaggle. An augmented data version of the COVID-19 dataset is also generated and analyzed to improve the quality of the solutions. Significant improvements are observed compared to state-of-the-art metaheuristic wrapper algorithms. We reported better classification results with feature selection than using the whole set of features of the dataset. 98.15% prediction accuracy is achieved with a 45% reduced number of features during experiments. We were successful in obtaining the new best solutions in the literature for the COVID-19 dataset.

INTRODUCTION

The CoronaVirus Disease 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) has become a big concern for all countries throughout the world (Albahri et al., 2020). The worldwide population has been devastated by the Coronavirus disease pandemic, which has overwhelmed advanced healthcare systems. As of February 2023, the COVID-19 pandemic has resulted in about 677 million cases and 6,784,798 deaths¹. Given the huge potential threat posed by the virus to public health, it is crucial to discover indicators that may be utilized as best predictors of COVID-19 patients' clinical outcomes and classify the severity of the patients as soon as possible (Lai et al., 2020; Dokeroglu et al., 2021). However, there is still a lack of sophisticated machine-learning applications in this field and this cannot be evaluated an easy task computationally. Considering their past achievements, wrapper classification algorithms that use metaheuristics can be very effective techniques in this domain.

Harris' Hawks Optimization (HHO) is a unique metaheuristic inspired by the collective behaviour of hawks. These avians use clever strategies, such as surprise pounce (seven kills), to grab their prey based on the escape tendencies of the victim. The HHO metaheuristic mimics the hawks' hunting habits to find the best/optimal solutions to the NP-Hard problems. In this study, we propose a novel parallel multi-objective HHO method (PHHO-KNN) using K-Nearest Neighbour (KNN) classifiers for predicting the mortality of COVID-19 patients according to their findings (features). Figure 1 gives a moment of a hawk's hunting scene.

Metaheuristic algorithms have to calculate many candidate solutions to reach the best result. Therefore, it becomes a difficult problem due to the size of the problem space. Parallel metaheuristics can be very effective tools when they are evaluated from this perspective. They can calculate much more fitness calculations than their single processor counterparts and obtain/achieve better results in shorter execution times. The diversity provided by many separate populations and the memory of different processors is

¹www.worldometers.info



Figure 1. The moment of a Harris hawk hunting a rabbit.

another advantage of these algorithms. Without getting stuck with local optima, they can explore and exploit the problem space in a better way (Alba et al., 2013). The algorithm we propose improves the accuracy of predictions using feature selection and augmented data as well. Feature selection is a critical field of machine learning Dokeroglu et al. (2022). The analysis of data, information retrieval, classification, and data mining all make use of feature selection intensively. It decreases the number of features by deleting data that is noisy, irrelevant, or redundant. This technique is an inevitable part of big data processing, which has become one of the most important problems of our world Bolón-Canedo et al. (2015). The KNN classification algorithm (Rajammal et al., 2022) is used as a classifier for each candidate solution generated by the HHO layer of the our proposed parallel algorithm.

We conduct comprehensive experiments on a recent real-world COVID-19 dataset from Kaggle. An augmented data version of the COVID-19 dataset is also generated to improve the quality of the solutions and analyzed. Significant improvements are observed compared to state-of-the-art metaheuristic wrapper algorithms. We were successful in obtaining the new best solutions for the COVID-19 dataset. Our contributions in this study can be listed as given below:

- A new parallel multi-objective HHO algorithm (PHHO-KNN) is proposed for the classification of COVID-19 patients.
- The results are improved using the augmented data version of the COVID-19 patients.
- Feature selection is applied to augmented COVID-19 data for the first time to the best of our knowledge.
- New best prediction accuracy results (in literature) are obtained with fewer features.
- Our proposed algorithm (PHHO-KNN) is the best-performing algorithm in the literature.

The second section of our paper reviews related studies of machine learning approaches for the COVID-19 disease. The third section gives details about our proposed parallel algorithm, PHHO-KNN. In the experimental setup and evaluation of results section, we define our experiments, software, datasets (including our proposed augmented dataset), hardware, and comparisons with state-of-the-art algorithms. In the last section, we give our concluding remarks and future work.

RELATED WORK

In this part, we give information about the recent state-of-the-art COVID-19 studies related to our work. Deniz et al. (2022) proposed a Multi-threaded evolutionary feature selection algorithm with Extreme Learning Machines (MG-ELM) to classify the severity of COVID-19 patients. They conduct experiments on the Kaggle dataset using feature selection. The results are competitive with other state-of-the-art algorithms. Dokeroglu et al. (2021) apply features selection and increase the accuracy level of predictions. They proposed new perching and besiege operators for the multi-objective classification problems. They conducted comprehensive experiments on UCI Machine Learning Repository. They applied the proposed algorithm to the Kaggle COVID-19 dataset. They report significant improvements compared to state-of-the-art metaheuristics. Dokeroglu et al. (2019) distinguished new and outstanding metaheuristics of the last two decades. They summarize the foundations of metaheuristics, recent trends, hybrid algorithms,

recent problems, parallel metaheuristics, and new opportunities. Dokeroglu and Sevinc (2019) proposed a Parallel Genetic ELM (IPE-ELM) for the data classification problem. The IPE-ELM uses feature selection and parallel fitness computation, parameter tuning of hidden neuron layers. The IPE-ELM is tested using UCI benchmark datasets. This algorithm has similar ideas to our proposed algorithm. Xue et al. (2018) presented a hybrid genetic algorithm and ELM (HGEFS). They proposed a novel mechanism to maintain diversity. Comparisons are carried out with benchmark datasets. The HGEFS outperforms other algorithms.

Kiziloz et al. (2018) proposed new multiobjective Teaching Learning Based Optimization (TLBO) algorithms for feature selection. They utilize the algorithm-specific parameterless concept of TLBO. Improvements are observed compared to NSGA-II, PSO, and Tabu Search algorithms. Dhamodharavadhani et al. (2020) studied Probabilistic, Radial Basis Function, and Generalized Regression Neural Networks for the mortality prediction of COVID-19 patients. The results showed that the Probabilistic and Radial Basis Function neural network performs better than other models. Cantú-Paz et al. (1998) presented a survey on parallel genetic algorithms (PGAs). They prepared a taxonomy of techniques and showed case studies. They also described the most significant issues in modelling and designing multi-population parallel GAs.

Bullock et al. (2020) presented an overview of Machine Learning studies to tackle many aspects of the COVID-19 crisis. We have identified applications and challenges posed by COVID-19. They reviewed datasets, tools, and resources needed to facilitate AI research. Chen et al. (2020) aim to develop a combination of feature selection algorithms (filter, wrapper, or embedded techniques). The experimental results showed that the combination of filter and wrapper algorithms is a better choice. Huang et al. (2010) studied ELM for classification and verified that the ELM tends to achieve better performance than SVM. The ELM is outstanding with its parameter settings. Huang et al. (2011) showed that the least square SVM and proximal SVM can be a unified framework and other regularization algorithms based to ELM can be built. Irshad et al. (2021) proposed a Particle Swarm Optimization algorithm (PSO) to obtain the optimal subset of retinal vessel features. They developed an objective function and compared the classification accuracy with the state-of-the-art approaches. The proposed algorithm is validated using a special dataset.

Iwendi et al. (2020) proposed a Random Forest model supported AdaBoost algorithm to predict the severity of the patients. The authors report 94% accuracy and 0.86 F1 Score. There is a positive correlation between the gender and deaths of the patients (in patients between the ages of 20 and 70). Kashef and Nezamabadi-pour (2015) introduced an Ant Colony Optimization (ABACO) feature selection algorithm. In this model, features are graph nodes and connected to other features. The result verified that the algorithm is good for feature selection. Mydukuri et al. (2022) proposed Gaussian neuro-fuzzy multi-layered data classification (LSRGNFM-LDC) method. The algorithm performs good prediction results on the 2019 COVID-19 dataset. Rasheed et al. (2020) surveyed state-of-the-art AI algorithms applied to the context of COVID-19 for mortality rates. Shaban et al. (2020) proposed a new hybrid feature selection method with KNN classifiers. Experiments showed that the strategy outperforms existing techniques.

Too and Mirjalili (2021) developed a novel Dragonfly Algorithm (HLBDA) for feature selection and data classification. The proposed method is applied to a coronavirus disease (COVID-19) dataset. The results demonstrated that the HLBDA has outstanding performance in accuracy and the number of features. Umarani and Subathra (2020) developed research using various machine learning algorithms and presented the application of data mining and machine learning techniques for COVID-19. Wu et al. (2020) developed a model for the severity and triage of COVID-19 patients. Yu et al. (2018) outlined recent AI studies and summarize the healthcare applications. Pizarro-Pennarolli et al. (2021) studied the impact of COVID-19 on daily activities they prepared a review. They studied post-infection cases of COVID-19 patients. Wang et al. (2020) proposed a random forest technique to predict the severity of COVID-19 patients. They applied a recursive feature elimination algorithm. Sheykhivand et al. (2021) introduced a new algorithm for the identification of COVID-19 using a deep neural network. Generative Adversarial Networks (GANs) are used with LSTM networks. They achieved 90% accuracy.

Shukla (2021) introduced a new gene selection technique with TLBO for cancer prediction. Experiments verified that the method can significantly remove the irrelevant genes and outperform the wrapper algorithms in terms of accuracy. Sun et al. (2020) proposed a feature selection with deep forest for the classification of COVID-19 chest tomography images. They use a deep forest model to learn a high-level

137 representation of the features. Experimental results with COVID-19 datasets showed that the proposed
138 algorithm achieves high performance in COVID-19. Tavarani (2020) prepared a comprehensive survey
139 on the applications of AI in battling against the difficulties the COVID-19 outbreak has caused. Yu et al.
140 (2020) proposed a genetic algorithm for the admission of vital cases. Several risk factors were identified
141 with clinical and statistical significance. They developed neural network models with selected variables
142 using Genetic Algorithms. The model outperformed generalized linear models.

143 PROPOSED PARALLEL MULTI-OBJECTIVE HARRIS HAWK ALGORITHM

144 There are two  in our proposed algorithm. In the first layer, HHO is used for feature selection
145 and KNN is employed for the classification of the patients. The hawks attack prey from different
146 directions cooperatively. The hawks use some patterns while hunting to exhaust the prey and make it
147 more vulnerable. The HHO is a gradient-free population-based method. The phases of the HHO are
148 inspired by the surprising jumps and attack strategies. Figure 2 shows the exploration and exploitation
149 steps of the HHO according to the energy and (q) values. In algorithm 1, the details of our PHHO-KNN
150 algorithm are presented.

151 The hawks watch the field to find prey by perching in trees. Perching happens according to the
152 locations of other hawks and the prey if the value of $q < 0.5$; otherwise, the hawks perch randomly. Two
153 exploration operators are proposed in this study. The exploration_1 chooses hawk_1 and hawk_2, and a
154 subset of features of hawk_1 is copied to hawk_2. The exploration_2 operator copies some selected features
155 of the best hawk to the current hawk. The newly generated candidate is inserted into the population. The
156 selection of the exploration or exploitation phase is decided with escaping energy value (E) of the prey.
157 The value (E) decreases as given below:

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

158 E_0 is the initial energy and T is the number of iterations. E_0 is between (-1, 1) and increases from
159 0 to 1 as the prey gets stronger. If E_0 decreases from 0 to -1, it means that the mobility of the prey is
160 decreasing. The value of E decreases with a higher number of iterations. If $|E| \geq 1$, the HHO performs
161 exploration; otherwise, it exploits.

162 The hawks can perform besiege by encircling the prey together according to the energy of the rabbit.
163 When $|E| \geq 0.5$ soft besiege, otherwise hard besiege is executed. When $r \geq 0.5$ and $|E| \geq 0.5$, it means the
164 rabbit's energy level is good, and it can perform random jumps to escape from the hawks (r is a random
165 value). The hawks can encircle the prey and use surprise jumps. A random number J , is used to create
166 these actions, and the features of the rabbit are copied to the current solution. If $r \geq 0.5$ and $|E| < 0.5$,
167 prey perform random jumps. The hawk's position is changed with the equation given below:

$$X(t+1) = X_{rabbit}(t) - E |\Delta X(t)| \quad (2)$$

168 $\Delta X(t)$ is the difference between the hawk and the current rabbit at iteration t . A single dimension of the
169 prey is copied (see Figure 3).

170
171 If $r < 0.5$ and $|E| \geq 0.5$, the rabbit can get rid of the attacks. A new operator (having a high perturbation
172 value) is proposed for this action. According to the prey's energy level, a subset of features is taken and
173 copied to the recent hawk.

174
175 When $|E| < 0.5$ and $r < 0.5$, the hawks are assumed to attack to reduce the distance. Some features
176 are chosen from the prey and copied to a randomly selected hawk. The minimum number of features
177 is used not to provide higher levels of perturbation. The pseudocode of the PHHO-KNN is
178 presented in Algorithm 1. Redundant candidate solutions are not allowed by the proposed algorithm.
179 Slave nodes/processors work independently from each other by generating their separate populations.
180 After the generations are finished, the slave nodes send their populations to the master node. The master
181 node receives the solutions and removes the similar solutions and reports the overall best 30 solutions that
182 construct a  to curve.

183 Figure 4 shows the representation of a solution. This structure denotes all the features of a solution.
184 Selected features are represented with ones and unselected ones are indicated as zeros.

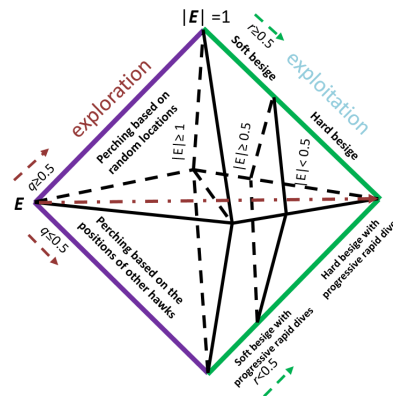


Figure 2. How the exploration and exploitation steps are decided by the HHO metaheuristic according to the energy level of the rabbit E , and random values q and r are given.

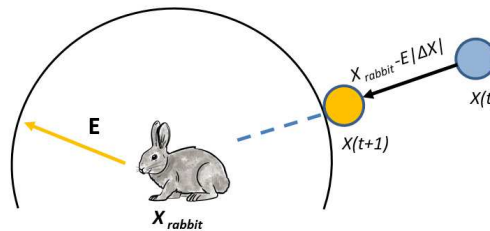


Figure 3. The positions of the vectors in hard besiege.

K-Nearest Neighbours Classifier (KNN)

We preferred KNN as a classifier in our algorithm because of its high performance in the recent classification studies (Rajammal et al., 2022). KNN is a non-parametric classification and regression technique used for supervised learning (Cunningham and Delany, 2021; Guo et al., 2003). The idea behind KNN is to identify the K closest data points to a given data point in the training set and then classify or predict the label of the new data point based on the labels of its K nearest neighbours. In KNN classification, the most common class among the K neighbours is assigned to the new data point. The choice of the value of K is an important hyperparameter in KNN, which can be selected using cross-validation. A small value of K may result in overfitting, while a large value of K may result in underfitting. KNN is a simple and effective algorithm that can work well with small and moderate-size datasets. It has been successfully applied to many applications, including image and speech recognition, text classification, and bioinformatics Batista et al. (2002); Larranaga et al. (2006). KNN can be computationally expensive, particularly when dealing with large datasets, and may require careful feature scaling to perform optimally. Figure 5 shows a case for the KNN classifier with $K=5$. Point X is a member of the green class because it has more neighbours from this set

We set the value of K as 5 for our proposed algorithm. Figure 6 shows the results of our experiments performed with K values from 1 to 70. The experiments are performed on 34 features of the original COVID-19 dataset with 5-fold cross-validation. Experiments with each K value are performed 30 times and average values are reported on the plot. The best values are obtained with a value of 5 for the K . How the performance of classification changes with different values of K can be observed easily with this experiment. Setting the value of K was a critical performance-increasing step for our study.

The Manhattan distance metric is used to measure the distance between two data points in our KNN classifier. It measures the distance between two points as if a taxi were driving through the grid-like street pattern of Manhattan. In contrast to Euclidean distance, which measures the straight-line distance between two points, Manhattan distance calculates the absolute differences between the corresponding coordinates of the two points and sums them up. It is often used in feature selection studies (Malkauthekar, 2013). We

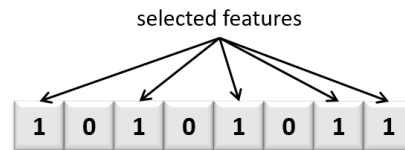


Figure 4. Representation of a set of selected features in a candidate solution.

Algorithm 1: Pseudocode of the parallel multi-objective HHO algorithm, PHHO-KNN

```

1 if (I am the master processor) then
2   while ( $i++ < \#slaves$ ) do
3     Receive solutions from slavei();
4     Update the set of best solutions();
5 if (I am a slave processor) then
6   Generate initial population of hawks  $X_i (i = 1, 2, \dots, N)$ 
7   while ( $i < max\_iter$ ) do
8     Find the best location and set it as the location of the rabbit ( $X_{rabbit}$ )
9     for (each hawk ( $X_i$ ) in the population) do
10      Update the energy level of the prey
11      if ( $Energy\_level \geq 1$ ) then
12        // Perform Exploration
13      else
14        // Perform Exploitation
15        if ( $Energy\_level \geq 0.5$ ) then
16          Perform soft besiege
17        else
18          Perform hard besiege
19      Calculate the fitness values of new hawks using KNN
20      Insert the hawks into the population
21  Send Results  $\leftarrow$  to the master node

```

211 have chosen the Manhattan metric because we observed that it produces better results in our experiments
 212 compared to Euclidean and Cosine similarity metrics.

213 EXPERIMENTAL SETUP AND EVALUATION OF THE RESULTS

214 In this part, we give the details of the experimental setup, information about the COVID-19 Dataset we
 215 use in our experiments, and the results and the comparison of our proposed algorithm to the state-of-
 216 the-art algorithms in the literature. Experiments are performed on an AMD Opteron Processor 6376.
 217 Its architecture is a Multiprocessing Non-Uniform Memory Access architecture. It has 4 sockets with
 218 eight cores each. A single core can support two threads. As a result, we have 64 CPUs to run 64 threads
 219 simultaneously. Each node is equipped with eight processors. These eight processors share a Last Level
 220 Cache of 6 MB. The system has 64 GB of RAM distributed across eight nodes. C++ programming
 221 language with MPI libraries is used during the implementation

222 COVID-19 Dataset

223 The original dataset is obtained from Kaggle ² and it has missing and redundant values. Therefore, we
 224 preprocessed this version of the dataset Iwendi et al. (2020); Dokeroglu et al. (2022). There are 1085 rows

²<https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset>

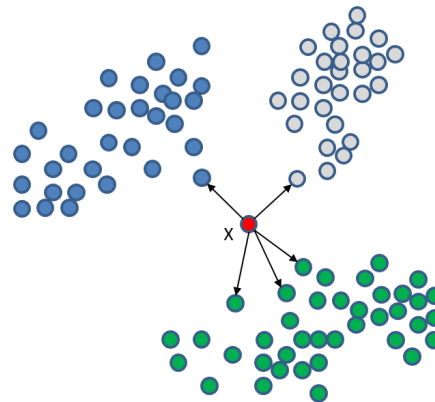


Figure 5. KNN classifier with $K=5$ for point X that is a member of the green class because it has more neighbours from this set.

(patients) with 34 features. Each symptom is defined as a new feature in this dataset. Integer encoding is employed for the text-based features.

We prepared our new augmented dataset using the original dataset of COVID-19 with 1085 rows. The new augmented dataset has 1192 rows (10% more than the original dataset). The new rows are selected randomly. No corrections are made to the classification data of the patients. With this new augmented dataset, 5% reductions are made the age of the patients, or updates were made on the data of the analysis results, not exceeding 1%. Our proposed algorithm is tested on both of these datasets. The following sections show how augmented data increases the prediction accuracy of our proposed algorithm.

The most informative features (from the COVID-19 dataset) selected by our proposed algorithm are age, diff_sym_hos, from_wuhan, location, hosp_vis, pneumonia, diff_sym_hos, fever, sym_on, vis_wuhan, diff_sym_hos, pneumonia, and difficulty in breathing.

Setting the best number of neighbours (K value) for the KNN classifier

In this part, we performed experiments with different values of K . Figure 6 shows the results of our experiments. Increasing the number of K up to 30 has good performance on the KNN classifier. Values larger than 50 have a negative effect on the performance of the KNN classifier for the COVID-19 dataset. We select the value of K as 5 because it gives us the highest accuracy results during our experiments. These results are the average values of 30 runs of values from 1 to 70. A difference up to 2.23% is observed between the best and worst results.

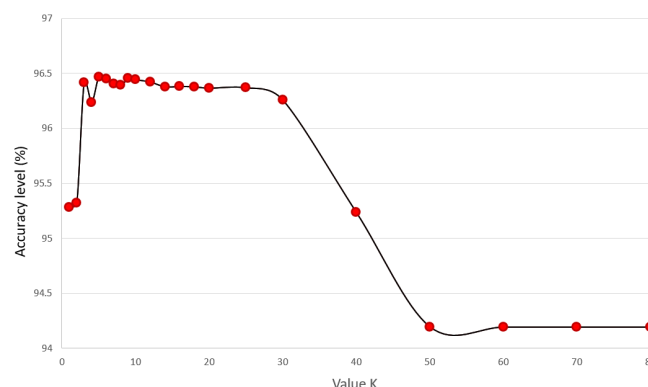


Figure 6. Experiments showing the effect of K value (the number of neighbours) on the prediction accuracy performance of the KNN classifier.

Experiments with augmented COVID-19 dataset

In this part, we use an augmented COVID-19 dataset in our experiments to improve the solution quality of the results. In data analysis, data augmentation techniques are used to expand the amount of data by adding slightly changed copies of the data or synthetic data (Kaushik et al. (2019)). When training a machine learning model, it functions as a regularizer and helps to minimize overfitting. It is strongly connected to data analysis oversampling. The size of the population is 30 and the number of generations is 500 (usually, 15.000 new candidate solutions are generated by the operators, and their fitness values are calculated during the optimization). When we consider the parallel computation environment with 64 processors, there are 64 different populations distributed to the memory of these processors and we perform 15.000 fitness evaluations for each population. This is a very huge effort for the classification of the patients in our dataset. Totally, 960.000 fitness values are calculated by our proposed algorithm. The initial populations are tried to breed different individuals by seeding the randomized functions with the id numbers of the processors.

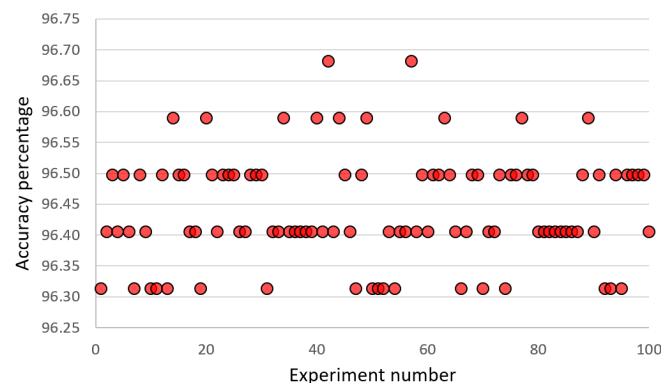


Figure 7. Results obtained with all features of the original COVID-19 datasets (100 samples are produced during the experiments).

Initial population and how it improves through generations

In this part, we compare the quality of our random initial population of the solutions with the final population quality. These experiments show that the initial population evolves after the execution of generations (iterations). The final population fits a better Pareto curve. Figures 8 and 9 show the comparison of the initial and final populations of the original COVID-19 and augmented COVID-19 datasets, respectively. The prediction accuracy results of the final populations are better in both cases for original and augmented versions of the COVID-19 datasets. In the original and augmented COVID-19 datasets, the accuracy levels are improved by 2.14% and 2.95% respectively (the average). Decreases in the number of features are two features for the augmented dataset while the average number of features is the same for the original COVID-19 dataset. These results are the average of 30 different executions of the experiments.

Figure 10 compares the quality of the last populations for the original and augmented datasets of the COVID-19 patients. It can be clearly seen that our proposed algorithm PHHO-KNN performs better with augmented data in terms of prediction accuracy (0.28% improvement is observed). For our second objective, the average number of features for original and augmented datasets are 17.5 and 18.63, respectively. Almost 45% reduction is achieved for the number of features we use during the classification.

Comparison with state-of-the-art algorithms

In this part, we compare our solutions with the state-of-the-art algorithms in the literature. Table 1 gives our comparisons to eight recent algorithms with the same COVID-19 dataset. All these algorithms are developed in the last two years. HLBDA (Too and Mirjalili, 2021), Boosted Random Forest (Iwendi et al., 2020), MHHO with ELM (Dokeroglu et al., 2021), LSRGNFM-LDC (Mydukuri et al., 2022), MG-ELM (Dokeroglu et al., 2021), MHHO with SVM (Dokeroglu et al., 2021), MHHO with Logistic Regression (Dokeroglu et al., 2021), and MHHO with Decision Trees (Dokeroglu et al., 2021) are the

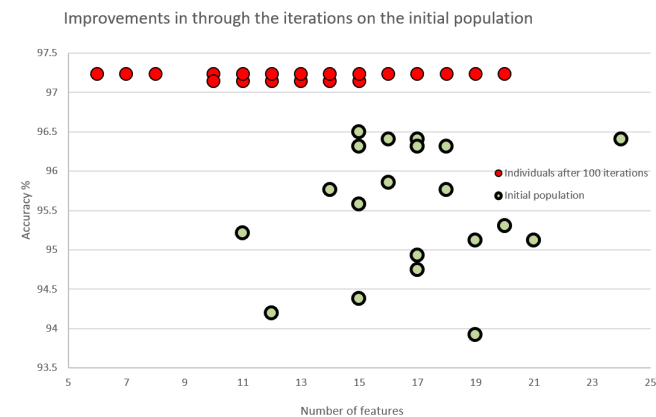


Figure 8. Comparison of the initial population with the evolved population after 15.000 iterations on the original COVID-19 dataset.

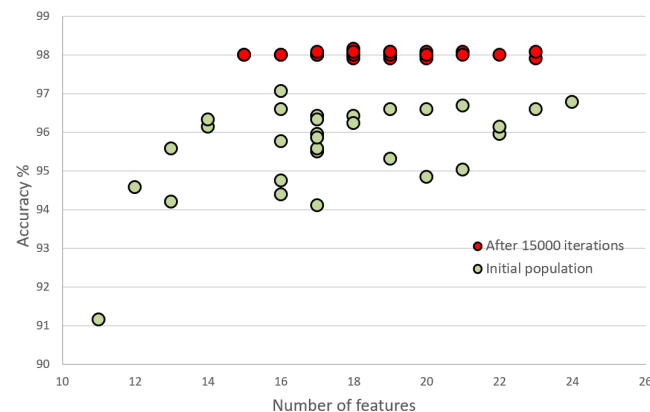


Figure 9. Comparison of the initial population with the evolved population after 15.000 iterations on the original augmented COVID-19 dataset.

best algorithms that have reported prediction accuracy results over 92.2%. Some of them are not multi-objective algorithms. We have obtained the results from their related studies. Our proposed algorithm PHHO-KNN outperforms all the algorithms in the literature. The MG-ELM algorithm is also a kind of parallel metaheuristic algorithm developed using Java threads. It was executed with 8 threads. The closest results to our solutions are provided by the MHHO algorithm with the Decision Trees algorithm. Our results are 0.20% and 0.47% better than this algorithm's solutions for the original COVID-19 and its augmented version, respectively.

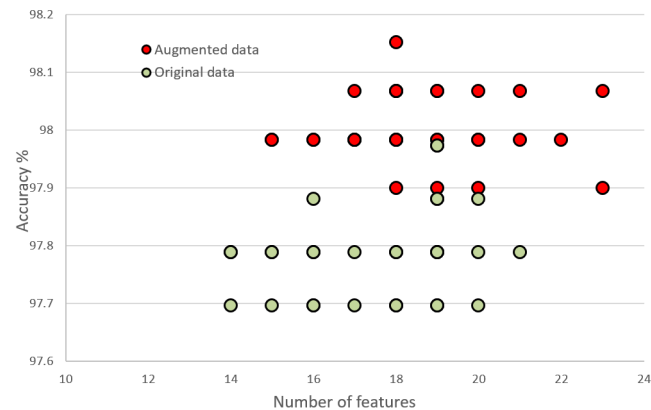


Figure 10. Comparing original COVID-19 dataset results with augmented dataset of COVID-19.

Table 1. Maximum accuracy values of the state-of-the-art algorithms on COVID-19 dataset. Sorted by their accuracy  es.

Method	Accuracy (%)
HLBDA (Too and Mirjalili, 2021)	92.21
Boosted Random Forest (Iwendi et al., 2020)	94.00
MHHO with ELM (Dokeroglu et al., 2021)	94.66
LSRGNFM-LDC (Mydukuri et al., 2022)	95.00
MG-ELM (Dokeroglu et al., 2021)	96.22
MHHO with SVM (Dokeroglu et al., 2021)	96.41
MHHO with Logistic Regression (Dokeroglu et al., 2021)	96.68
MHHO with Decision Trees (Dokeroglu et al., 2021)	97.61
PHHO-KNN (our algorithm)	97.88
PHHO-KNN (with augmented data)	98.15

CONCLUSIONS AND FUTURE WORK

This study proposed a new robust parallel multi-objective HHO algorithm (PHHO-KNN) for the mortality prediction of COVID-19 patients. Parallel metaheuristics are very powerful tools with their scalable abilities to optimize NP-hard problems effectively. The new metaheuristic HHO is used for the feature selection layer of the algorithm and KNN is used as a classifier for the COVID-19 dataset. The PHHO-KNN algorithm is adaptable and can be applied to the classification of any other dataset easily. It has been observed that the number of fitness calculations increases depending on the number of processors in the computation environment and that the results can be improved as the number of added processors increases. When we compare the results of our algorithm with state-of-the-art studies in literature, we can say that our results constitute the new best results in this domain. With the augmented dataset, the solution quality is increased from 97.88% to 98.15% in terms of accuracy. Almost 45% reduction is achieved for the number of features in the average.

In future, we intend to employ deep learning classifiers with our parallel multi-objective HHO algorithm. Hyper-heuristics is an attractive field of research. Applying state-of-the-art hyper-heuristic frameworks to this domain can be an impressive good contribution. Moreover, we are considering examining the newly published COVID datasets.

REFERENCES

- Alba, E., Luque, G., and Nesmachnow, S. (2013). Parallel metaheuristics: recent advances and new trends. *International Transactions in Operational Research*, 20(1):1–48.
- Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-Qays, Z., Zaidan, A., Zaidan, B., Albahri, A., AlAmoodi, A. H., Khlaf, J. M., Almahdi, E., et al. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (covid-19): a systematic review. *Journal of medical systems*, 44:1–11.
- Batista, G. E., Monard, M. C., et al. (2002). A study of k-nearest neighbour as an imputation method. *His*, 87(251-260):48.
- Bolón-Canedo, V., Sánchez-Marono, N., and Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-based systems*, 86:33–45.
- Bullock, J., Luccioni, A., Pham, K. H., Lam, C. S. N., and Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against covid-19. *Journal of Artificial Intelligence Research*, 69:807–845.
- Cantú-Paz, E. et al. (1998). A survey of parallel genetic algorithms. *Calculateurs paralleles, reseaux et systems repartis*, 10(2):141–171.
- Chen, C.-W., Tsai, Y.-H., Chang, F.-R., and Lin, W.-C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, 37(5):e12553.
- Cunningham, P. and Delany, S. J. (2021). k-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR)*, 54(6):1–25.
- Deniz, A., Kiziloz, H. E., Sevinc, E., and Dokeroglu, T. (2022). Predicting the severity of covid-19 patients using a multi-threaded evolutionary feature selection algorithm. *Expert Systems*, 39(5):e12949.
- Dhamodharavadhani, S., Rathipriya, R., and Chatterjee, J. M. (2020). Covid-19 mortality rate prediction for india using statistical neural network models. *Frontiers in Public Health*, 8:441.
- Dokeroglu, T., Deniz, A., and Kiziloz, H. E. (2021). A robust multiobjective harris’ hawks optimization algorithm for the binary classification problem. *Knowledge-Based Systems*, 227:107219.
- Dokeroglu, T., Deniz, A., and Kiziloz, H. E. (2022). A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing*.
- Dokeroglu, T. and Sevinc, E. (2019). Evolutionary parallel extreme learning machines for the data classification problem. *Computers & Industrial Engineering*, 130:237–249.
- Dokeroglu, T., Sevinc, E., Kucukyilmaz, T., and Cosar, A. (2019). A survey on new generation metaheuristic algorithms. *Computers & Industrial Engineering*, 137:106040.
- Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer.

- 339 Huang, G.-B., Ding, X., and Zhou, H. (2010). Optimization method based extreme learning machine for
340 classification. *Neurocomputing*, 74(1-3):155–163.
- 341 Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2011). Extreme learning machine for regression and
342 multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*,
343 42(2):513–529.
- 344 Irshad, S., Yin, X., and Zhang, Y. (2021). A new approach for retinal vessel differentiation using
345 binary particle swarm optimization. *Computer Methods in Biomechanics and Biomedical Engineering:
346 Imaging & Visualization*, 9(5):510–522.
- 347 Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S.,
348 and Jo, O. (2020). Covid-19 patient health prediction using boosted random forest algorithm. *Frontiers
349 in public health*, 8:357.
- 350 Kashef, S. and Nezamabadi-pour, H. (2015). An advanced aco algorithm for feature subset selection.
351 *Neurocomputing*, 147:271–279.
- 352 Kaushik, D., Hovy, E., and Lipton, Z. C. (2019). Learning the difference that makes a difference with
353 counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.
- 354 Kiziloz, H. E., Deniz, A., Dokeroglu, T., and Cosar, A. (2018). Novel multiobjective tlbo algorithms for
355 the feature subset selection problem. *Neurocomputing*, 306:94–107.
- 356 Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J., and Hsueh, P.-R. (2020). Severe acute respiratory syndrome
357 coronavirus 2 (sars-cov-2) and coronavirus disease-2019 (covid-19): The epidemic and the challenges.
358 *International journal of antimicrobial agents*, 55(3):105924.
- 359 Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armananzas, R.,
360 Santafé, G., Pérez, A., et al. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*,
361 7(1):86–112.
- 362 Malkauthekar, M. (2013). Analysis of euclidean distance and manhattan distance measure in face recogni-
363 tion. In *Third International Conference on Computational Intelligence and Information Technology
364 (CIIT 2013)*, pages 503–507. IET.
- 365 Mydukuri, R. V., Kallam, S., Patan, R., Al-Turjman, F., and Ramachandran, M. (2022). Deming least
366 square regressed feature selection and gaussian neuro-fuzzy multi-layered data classifier for early covid
367 prediction. *Expert Systems*, 39(4):e12694.
- 368 Pizarro-Pennarolli, C., Sánchez-Rojas, C., Torres-Castro, R., Vera-Urbe, R., Sanchez-Ramirez, D. C.,
369 Vasconcello-Castillo, L., Solís-Navarro, L., and Rivera-Lillo, G. (2021). Assessment of activities of
370 daily living in patients post covid-19: a systematic review. *PeerJ*, 9:e11026.
- 371 Rajammal, R. R., Mirjalili, S., Ekambaram, G., and Palanisamy, N. (2022). Binary grey wolf optimizer
372 with mutation and adaptive k-nearest neighbour for feature selection in parkinson’s disease diagnosis.
373 *Knowledge-Based Systems*, 246:108701.
- 374 Rasheed, J., Jamil, A., Hameed, A. A., Aftab, U., Aftab, J., Shah, S. A., and Draheim, D. (2020). A
375 survey on artificial intelligence approaches in supporting frontline workers and decision makers for the
376 covid-19 pandemic. *Chaos, Solitons & Fractals*, 141:110337.
- 377 Shaban, W. M., Rabie, A. H., Saleh, A. I., and Abo-Elsoud, M. (2020). A new covid-19 patients detection
378 strategy (cpds) based on hybrid feature selection and enhanced knn classifier. *Knowledge-Based
379 Systems*, 205:106270.
- 380 Sheykhivand, S., Mousavi, Z., Mojtahedi, S., Rezaii, T. Y., Farzamnia, A., Meshgini, S., and Saad, I.
381 (2021). Developing an efficient deep neural network for automatic detection of covid-19 using chest
382 x-ray images. *Alexandria Engineering Journal*, 60(3):2885–2903.
- 383 Shukla, A. K. (2021). Feature selection inspired by human intelligence for improving classification
384 accuracy of cancer types. *Computational Intelligence*, 37(4):1571–1598.
- 385 Sun, L., Mo, Z., Yan, F., Xia, L., Shan, F., Ding, Z., Song, B., Gao, W., Shao, W., Shi, F., et al. (2020).
386 Adaptive feature selection guided deep forest for covid-19 classification with chest ct. *IEEE Journal of
387 Biomedical and Health Informatics*, 24(10):2798–2805.
- 388 Tayarani, M. (2020). Applications of artificial intelligence in battling against covid-19: A literature review.
389 *Chaos, Solitons & Fractals*.
- 390 Too, J. and Mirjalili, S. (2021). A hyper learning binary dragonfly algorithm for feature selection: A
391 covid-19 case study. *Knowledge-Based Systems*, 212:106553.
- 392 Umarani, V. and Subathra, M. (2020). Data mining and machine learning techniques in prediction of
393 covid-19 outbreaks-a recent review. *Tierärztliche Praxis*, 40:1437–1447.

- 394 Wang, J., Yu, H., Hua, Q., Jing, S., Liu, Z., Peng, X., Luo, Y., et al. (2020). A descriptive study of random
395 forest algorithm for predicting covid-19 patients outcome. *PeerJ*, 8:e9945.
- 396 Wu, G., Yang, P., Xie, Y., Woodruff, H. C., Rao, X., Guiot, J., Frix, A.-N., Louis, R., Moutschen, M.,
397 Li, J., et al. (2020). Development of a clinical decision support system for severity risk prediction
398 and triage of covid-19 patients at hospital admission: an international multicentre study. *European*
399 *Respiratory Journal*, 56(2).
- 400 Xue, X., Yao, M., and Wu, Z. (2018). A novel ensemble-based wrapper method for feature selection using
401 extreme learning machine and genetic algorithm. *Knowledge and Information Systems*, 57:389–412.
- 402 Yu, K.-H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical*
403 *engineering*, 2(10):719–731.
- 404 Yu, Y., Zhu, C., Yang, L., Dong, H., Wang, R., Ni, H., Chen, E., and Zhang, Z. (2020). Identification of
405 risk factors for mortality associated with covid-19. *PeerJ*, 8:e9885.