

# A Modified Whale Optimization Algorithm for Enhancing the Features Selection Process in Machine Learning

Ezaz Uddin Syed  
Electrical Engineer  
AGES Consultants Ltd  
Peshawar, Pakistan  
ezaz.usyed@gmail.com

Mohsin Masood  
Department of Computing &  
Technology  
Abasyn University Pakistan  
mohsin.masood@abasyn.edu.pk

Mohamed Mostafa Fouad  
Arab Academy for Science,  
Technology, Maritime, Transport  
Cairo, Egypt  
mohamed\_mostafa@aast.edu

Ivan Glesk  
Electronics and Electrical  
Department  
Glasgow, UK  
ivan.glesk@strath.ac.uk

**Abstract** – In recent years, when there is an abundance of large datasets in various fields, the importance of feature selection problem has become critical for researchers. The real world applications rely on large datasets, which implies that datasets have hundreds of instances and attributes. Finding a better way of optimum feature selection could significantly improve the machine learning predictions. Recently, metaheuristics have gained momentous popularity for solving feature selection problem. Whale Optimization Algorithm has gained significant attention by the researcher community searching to solve the feature selection problem. However, the exploration problem in whale optimization algorithm still exists and remains to be researched as various parameters within the whale algorithm have been ignored. This paper proposes a new and improved version of the whale algorithm entitled Modified Whale Optimization Algorithm (MWOA) that hybrid with the machine learning models such as logistic regression, decision tree, random forest, K-nearest neighbor, support vector machine, naïve Bayes model. To test this new approach and the performance, the breast cancer dataset was used for MWOA evaluation. The test results revealed the superiority of this model when compared to the results obtained by machine learning models.

**Index Terms** – Artificial intelligence, dimension reduction, feature selection problem, machine learning Models, whale optimization algorithm.

## I. INTRODUCTION

Artificial intelligence techniques are progressively becoming an essential part of the research process in many branches of science. Searching within huge amount of data becomes increasingly problematic particularly when searching for features that could be used for classification and prediction of problems. Feature selection is a method that initiates steps to recognize the main attributes of the provided dataset. Feature selection methods are applied to a range of biological, finance, and intrusion detection systems problems. Out of many applications where a feature selection has proved successful is medicine. It helped in easing the dimensionality and analyses of the causes of diseases. During the data analyses, there are often found features in the datasets that are irrelevant and should be ignored. Feature selection allows filtering these irrelevant features and only relevant features are selected. Various methodologies have been proposed to handle the situation. As an example, random search, greedy search and many more have been used as techniques to select ideal subsets of features. These methods face drawbacks from untimely convergence, from extreme complexity, and extensive computation. Researchers have sought of metaheuristic

algorithms to deal with these aforementioned challenges. There are number of ways to solve these problems. Metaheuristic algorithms have proven to provide better results than other procedures. [1-3].

To acquire the best subset of features, several selection approaches have been developed such as filter, wrapper, and embedding approaches. The classification algorithms are always included in wrapper approaches, and they interact with the classifier. These approaches are more time consuming than filters, but they also produce more ideal results when equated to filter approaches. Filters and wrapper approaches are combined to form embedded approaches. The feature selection is a part of the training process in embedded approaches, and the classifier is used in the training process. [1-3]. If compared to filter approaches, wrapper approaches produce superior outcomes, but they are slower. Wrapper approaches rely on the modelling algorithm, which generates and evaluates each subset. It uses a different search strategy to generate subsets. Randomness is used to explore the search space in randomised algorithms, which prevents them from being held in local optima. Population-based techniques, such as simulated annealing, random generation, and metaheuristic algorithms, are examples of randomised algorithms. [5-9]

Machine learning methods on the other hand are seen to solve the feature selection problem but with a great cost in computation, high complexities and premature convergence. In order to contain these drawbacks, metaheuristic algorithms were proposed, as they are good at dealing with these type of conditions. Metaheuristic algorithms are used with classifiers to solve feature selection problems [5-9].

The paper has arranged in the form of sections. Section II presents a comprehensive literature review of various versions of whale optimization algorithm for solving feature selection problem. The proposed algorithm is briefly explained in Section III in the form of pseudocode. Experimental setup of the proposed methodology along with results are explained in Section IV, whereas Conclusion is given in Section V.

## II. LITERATURE REVIEW

Metaheuristic algorithms begin their optimization process by producing random solutions. Unlike gradient search approaches, it does not necessitate calculating the derivative of the search space. With the ability of simple principle and easy implementation, metaheuristic algorithms are versatile and straightforward. Easy modifications can be made to suit the

target problem. The major feature of metaheuristic algorithms is their capability to avoid algorithms from converging prematurely. Because algorithms have stochastic behaviour, they act as a black box, avoiding local optima and efficiently and effectively exploring the search space. The algorithms strike a balance between exploration and exploitation, which are the two most important components of the system [10]. They have been successfully applied to a varied range of engineering and science problems, including those in electrical fields (power generation), industrial (scheduling jobs, transport, routing problem), civil engineering (designing structures), communication (designing networks and radars), and data mining (classification, prediction).

Whale Optimization Algorithm “WOA” is based on the special behaviour of the hunting method of humpback whales called bubble-net feeding method. What they do is use two manoeuvres for foraging. This method of bubble-net is then mathematically modelled in order to perform optimization [11]. Hussien et al. [12-13] proposed binary variant of the basic whale optimization algorithm as a method for solving the feature selection problem. Sayed et al. [14] proposes “WOA” algorithm using chaotic properties, such as regularity and semi-stochastic and hence the “WOA” algorithm is improved by comparing ten chaotic maps. To reduce the search space explored by “WOA”, Tubishat et al. [15] used SVM classifier along with information gain as filter features selection technique with “WOA”. Mafarja and Mirjalili [16] proposed improvements to the native “WOA” by applying Tournament and Roulette Wheel selection mechanisms WOA-T and WOA-R. Another approach for improvement was crossover and mutation operators WOA-CM. Results demonstrate the efficiency of the proposed approaches in finding the optimal feature subsets. A quantum-based version of Whale Optimization Algorithm (QWOA) presented by Agrawal et al. [17] made improvements to the convergence and diversification properties of the classical Whale Optimization Algorithm for feature selection. Modified mutation and crossover operators are also used in the techniques of whales in the proposed QWOA. Zheng et al. [18] proposed a hybrid feature selection algorithm named maximum Pearson maximum distance improved whale optimization algorithm (MPMDIWOA). The algorithm involves maximum Pearson maximum distance (MPMD) and improved whale optimization algorithm (IWOA), maximum value without change (MVWC) and threshold changes to adjust the order and frequency of execution. Bui et al. [19] proposed a hybrid makeover of Whale Optimization Algorithm and Adaptive Neuro-Fuzzy Inference System (ANFIS). The algorithm was for feature reduction and classification of segmented objects. Results showed an improvement in image classification accuracy. Tawhid et al. [20] propose an optimal feature selection algorithm based on binary whale optimization (BWOA) for an efficient search approach that is able to efficiently find the minimal feature subset.

Hybridization of metaheuristics with machine learning models (ML) is recognized an attractive approach for achieving optimized ML models for various datasets. The dominant argument of using metaheuristics for ML models optimization is the presence of large datasets, which are based on hundreds

of features. To optimize the model, the optimal features need to be selected, known as feature selection problem. This problem is considered as non-polynomial (NP) hard optimization problem. Therefore, number of attempts have been made in the form of research, as thoroughly discussed in literature. In addition to this, whale optimization algorithms suffer from deficiencies like dimensionality reduction and classification problem as identified in [12-13]; entrapment in local optima and slow convergence speed [14]; feature selection problems [15], [20]; exploration & exploitation problems [16], [17]; optimal classification accuracy [18], [19]. Therefore, there is a need to offer an improved version of “WOA” that could target one of the mentioned problems. This paper offers a modified version of “WOA” that could overcome “WOA” algorithm’s local optima problem. It targets the features selection problem and generates improved solutions.

Algorithm 1: Modified Whale Optimization Algorithm

---

**Output:** Features Selection

```

initialize whales population ( $x_i, i = 1,2,3, \dots n$ )
initialize whales position, range [0,1]
  if position value > 1 then
    | feature selected
  else
    | feature is not selected
compute fitness function ( $ff$ ) using ML models
find set of features with high accuracy
While maximum iteration < iteration
  for each whale in the swarm
    update  $\vec{A}, \vec{C}, l, p$ 
    if  $p < 0.5$ 
      if  $|A| < 1$ 
        | update whale positions
      else if  $|A| \geq 1$ 
        | select random whale position
        | update whale positions
      end
    else if  $p \geq 0.5$ 
      | update whale positions
    end
    select features using positions
    compute fitness function using ML models
    if  $ff \leq \text{previous } ff$  (not optimal)
      | choose features from memory
      | update positions mapped with features
    else
      | compute position, range [0,1] for features
      | compute the best set of features
      | maintain memory for selected features
    end
  end
end

```

---

### III. PROPOSED METHOD

Although the “WOA” also derived some of its structures from different nature-inspired algorithms, integration with other algorithms or modification with new parameters needs to be researched more to upsurge its versatility. Furthermore,

Accuracy Comparison between ML Models V/s Optimized ML with Modified- WOA Model

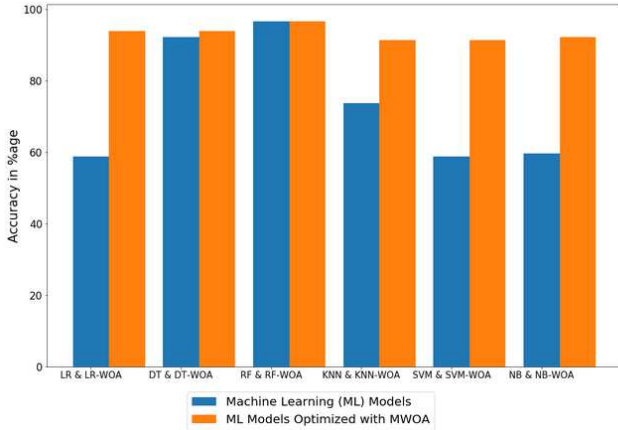


Fig. 1: Accuracy comparison of ML and MWOA-ML models

Sensitivity Comparison between ML Models V/s Optimized ML with Modified- WOA Model

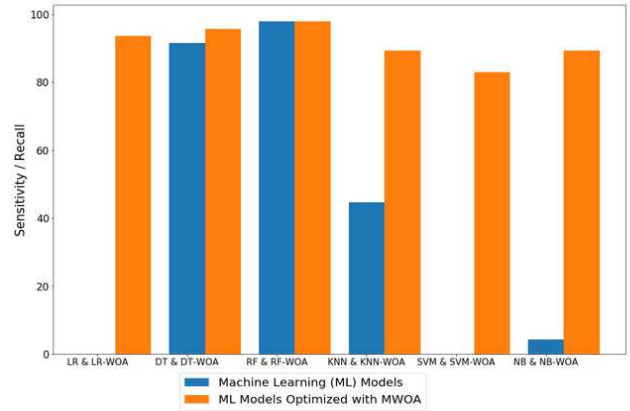


Fig. 2: Sensitivity of ML and MWOA-ML models

Classification Error Comparison between ML Models V/s Optimized ML with Modified- WOA Model

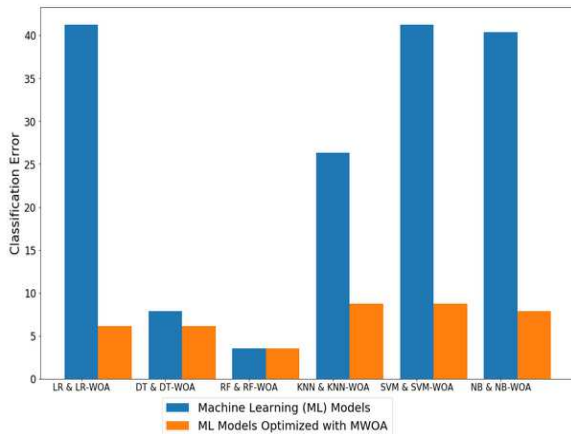


Fig. 3: Classification Error of ML and MWOA-ML models

Precision Comparison between ML Models V/s Optimized ML with Modified- WOA Model

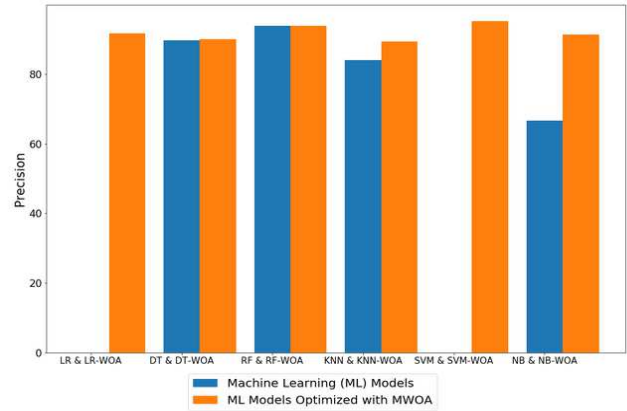


Fig. 4: Precision of ML and MWOA-ML models

hybridization of “WOA” with ML models attain their boundary if significant problem instances with large search spaces become achievable solutions. Therefore, this paper proposed a modified version of “WOA”, titled as “*Modified Whale Optimization Algorithm (MWOA)*”. This recommended algorithm is hybridized with prominent ML models *i. e.* logistic regression (LR), decision tree (DT), random forest (RF), K-nearest neighbours (KNN), support vector machine (SVM), naïve Bayes (NB) model. The proposed MWOA algorithm with ML is further explained in the form of pseudocode (Algorithm 1), where modification in the original “WOA” is highlighted with blue colour.

#### IV. EXPERIMENTAL RESULTS

The experimental work conducted in this paper is simulated over Python tool using Pandas, Numpy, Scikit-learn, Matplot, and Seaborn libraries. The models are tested over Wisconsin Breast cancer dataset, which is based on 32 features and 569 instances. This dataset is used for the classification as for binary classification models. The experiments are conducted in such a way that each model (LR, DT, RF, KNN, SVM, NB) is tested over given dataset. With the same dataset, the proposed MWOA hybrid with all models are tested and compared with the

classification performance parameters such as accuracy rate, classification error, sensitivity, precision. The results are plotted in the form of bar charts (*Fig. 1, Fig. 2, Fig. 3, and Fig. 4*) for make a comprehensive analysis of the obtained results for each ML model with its competitor proposed MWOA hybrid ML models. For clarity, different colours are used in bar plotted results, such as blue for ML models and orange for MWOA-hybrid with ML models.

The results in *Fig. 1* depicts the accuracy rate of each model (LR, DT, RF, KNN, SVM, NB) in comparison with proposed MWOA-hybrid ML Models (LR-WOA, DT-WOA, RF-WOA, KNN-WOA, SVM-WOA, NB-WOA). Accuracy identifies the overall correction rate of ML model that how often the model successfully predicts or classify the instances by having certain number of features. When analysing the results in *Fig.1*, it is revealed the all ML models that were hybridized with proposed MWOA has offered an improved results compare to their counterpart ML models. For instance, a major improvement of model accuracy can be noticed in LR-WOA, KNN-WOA, SVM-WOA, and NB-WOA compare to LR, KNN, SVM, and NB ML models. However, DT, and RF have already obtained their maximum accuracy and hence, have either same or very close accuracy with DT-WOA, and RF-WOA models.

*Fig. 2* illustrates the results of sensitivity parameter for each ML model with its counterpart MWOA-ML models. According to *Fig. 2* results, LR, SVM, and NB has very poor classification results compare to LR-WOA, SVM-WOA, and NB-WOA models. For example, LR, SVM and NB totally failed to classify the individuals having disease in the dataset. These bar graphs for LR, SVM, and NB also revealed that these models should not be used for this type of medical datasets.


Classification error is another important performance parameter for classification models that is used to highlight the incorrect classification of ML models. The results in *Fig. 3* shows that ML models are failed abundant times to make right predictions of 32 features based dataset. On the other hand, the proposed model offered much better-quality results of classification-error. LR, KNN, SVM, and NB are dramatically tinted with their poor classifications. For instance, LR, KNN, SVM, and NB models have more than 40% classification error, which revealed that these models failed to classify the individuals as not patients in the dataset.

*Fig. 4* shows the precision of each ML model and compared with its opponent MWOA-hybrid ML models in the form of bar graph. Precision parameters identifies only YES cases, such that when the instance or individual actually have breast cancer disease and the model also predicted right as YES for the given instance. The results in *Fig. 4* exposed the poor classification of LR and SVM, as these two models failed to classify the YES cases of the individuals of the dataset. On the contrary, the proposed MWOA-hybrid with ML models have outclass its counterpart models (LR, and SVM) and also have impressive results for the considered breast cancer dataset.

## V. CONCLUSION

The research presented in this paper offers a novel approach called the *Modified Whale Optimization Algorithm (MWOA)* for solving the feature selection problem in machine learning models. To analyze the model's performance, a number of performance parameters were considered such as: accuracy rate, classification error, precision, and sensitivity. The results obtained are then compared with machine learning classification models. As documented, the proposed MWOA-ML models offer a major improvement in the form of the accuracy, classification-error, sensitivity, and precision. These results also help to expose the poor performance of LR, SVM and NB models. This confirms that these models failed to offer better prediction results for the given dataset. In summary, this paper presents an approach that improves ML models with prediction in the field of "medical modeling". It is also noted that algorithm presented limited improved results for random forest and decision tree models. Therefore, there is a need to do further modification in the presented technique for the mentioned machine learning models.

## ACKNOWLEDGMENT

This research is funded by Roryan Pharmaceutical Ltd with the Grant ID: RPI/EXP-127 and the European Union Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 734331. 

- [1] Remeseiro, B. and Bolon-Canedo, V., 2019. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112, p.103375.
- [2] Masood, M., Fouad, M. and Glesk, I., 2018. Analysis of Artificial Intelligence-Based Metaheuristic Algorithm for MPLS Network Optimization. 2018 20th International Conference on Transparent Optical Networks (ICTON).
- [3] Jan, S. and Masood, M., 2021. Multiple Solutions Based Particle Swarm Optimization for Cluster-Head-Selection in Wireless-Sensor-Network. 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2).
- [4] Masood, M., Fouad, M. and Glesk, I., 2018. Pareto Based Bat Algorithm for Multi Objectives Multiple Constraints Optimization in GMPLS Networks. *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, pp.33-41.
- [5] Agrawal, P., Abutarboush, H., Ganesh, T. and Mohamed, A., 2021. Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019). *IEEE Access*, 9, pp.26766-26791.
- [6] Masood, M., Fouad, M. and Glesk, I., 2017. A Pareto based approach with elitist learning strategy for MPLS/GMPS networks. 2017 9th Computer Science and Electronic Engineering (CEECE).
- [7] Masood, M., Fouad, M. and Glesk, I., 2017. Proposing bat inspired heuristic algorithm for the optimization of GMPLS networks. 2017 25th Telecommunication Forum (TELFOR).
- [8] Masood, M., Fouad, M., Kamal, R., Glesk, I. and Khan, I., 2019. An Improved Particle Swarm Algorithm for Multi-Objectives Based Optimization in MPLS/GMPLS Networks. *IEEE Access*, 7, pp.137147-137162.
- [9] Masood, M., Fouad, M., Seyedzadeh, S. and Glesk, I., 2019. Energy Efficient Software Defined Networking Algorithm for Wireless Sensor Networks. *Transportation Research Procedia*, 40, pp.1481-1488.
- [10] Hussain, K., Salleh, M., Cheng, S. and Shi, Y., 2018. On the exploration and exploitation in popular swarm-based metaheuristic algorithms. *Neural Computing and Applications*, 31(11), pp.7665-7683.
- [11] Mirjalili, S. and Lewis, A., 2016. The Whale Optimization Algorithm. *Advances in Engineering Software*, 95, pp.51-67.
- [12] Hussien, A., Hassanien, A., Houssein, E., Bhattacharyya, S. and Amin, M., 2018. S-shaped Binary Whale Optimization Algorithm for Feature Selection. *Recent Trends in Signal and Image Processing*, pp.79-87.
- [13] Hussien, A., Houssein, E. and Hassanien, A., 2017. A binary whale optimization algorithm with hyperbolic tangent fitness function for feature selection. 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS).
- [14] Sayed, G., Darwish, A. and Hassanien, A., 2018. A New Chaotic Whale Optimization Algorithm for Features Selection. *Journal of Classification*, 35(2), pp.300-344.
- [15] Tubishat, M., Abushariah, M., Idris, N. and Aljarah, I., 2018. Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. *Applied Intelligence*, 49(5), pp.1688-1707.
- [16] Mafarja, M. and Mirjalili, S., 2018. Whale optimization approaches for wrapper feature selection. *Applied Soft Computing*, 62, pp.441-453.
- [17] Agrawal, R., Kaur, B. and Sharma, S., 2020. Quantum based Whale Optimization Algorithm for wrapper feature selection. *Applied Soft Computing*, 89, p.106092.
- [18] Zheng, Y., Li, Y., Wang, G., Chen, Y., Xu, Q., Fan, J. and Cui, X., 2019. A Novel Hybrid Algorithm for Feature Selection Based on Whale Optimization Algorithm. *IEEE Access*, 7, pp.14908-14923.
- [19] Bui, Q., Pham, M., Nguyen, Q., Nguyen, L. and Pham, H., 2019. Whale Optimization Algorithm and Adaptive Neuro-Fuzzy Inference System: a hybrid method for feature selection and land pattern classification. *International Journal of Remote Sensing*, 40(13), pp.5078-5093.
- [20] Tawhid, M. and Ibrahim, A., 2019. Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm. *International Journal of Machine Learning and Cybernetics*, 11(3), pp.573-602.