QUBAHAN ACADEMIC JOURNAL
VOL. 4, NO. 1, March 2024
https://doi.org/10.48161/qaj.v4n1a466

Improved Ozone Level Detection through Feature Selection with Modified Whale Optimization Algorithm

Li Yu Yab [1], Noorhaniza Wahid [1] and Rahayu A. Hamid [1]

[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Johor, Malaysia;

Corresponding author: e-mail: nhaniza@uthm.edu.my.

ABSTRACT: This study presents a new approach for ozone level detection through feature selection by the modified Whale Optimization Algorithm (mWOA). This study aims to enhance the accuracy and efficiency of ozone level prediction models by selecting the most informative features from the dataset. As air quality deterioration poses significant risks to both human health and ecological equilibrium, pinpointing relevant features becomes essential for boosting prediction accuracy. The scope of the research includes comparing the performance of mWOA with the original WOA in two feature selection techniques: filter-based and wrapper-based. The experiments run proposed approaches on a multivariate time-series dataset with 20 repetitions. The evaluation criteria include processing time, number of features selected, and classification accuracy obtained by the kNN classifier. The statistical results demonstrate the effectiveness of the proposed mWOA approach, outperforming WOA due to the modified control parameter that enables a more precise exploration of the search area. The findings of this study reveal the improved performance of mWOA in selecting informative features, resulting in better prediction on average: 93.75% for filter-based and 94.49% for wrapper-based. In conclusion, the wrapper-based feature selection using the mWOA approach proves to be a valuable asset in enhancing the accuracy and efficiency of ozone level detection models. In the future, the proposed technique can be used for more applications in environmental science and engineering research.

Keywords: ozone level detection, whale optimization algorithm, inversed control parameter, feature selection, k-nearest neighbours.

I. INTRODUCTION

The Earth has a complex atmosphere system that plays a crucial role in supporting life on our planet. Among its many components, ozone (O3) stands out as a critical gas with both beneficial and detrimental effects [1]. Ozone is found in different layers of the atmosphere, including the ozone layer in the stratosphere, where it acts as a shield against harmful ultraviolet radiation [2]. Under the influence of sunlight, certain pollutants, including nitrogen oxides (NOx) and volatile organic compounds (VOCs) undergo chemical reactions, forming ground-level ozone [3-5]. This process typically happens in urban areas with high traffic and industrial activities [1-5]. Notably, ozone levels in Malaysia are closely linked to transportation-related emissions, as vehicular pollutants contribute significantly to the chemical reactions that lead to elevated ozone concentrations in the atmosphere [6, 7].

At ground level, ozone can become a major air pollutant, posing serious health risks to humans and ecosystems [1-8]. Ground-level ozone can also irritate the respiratory system, leading to respiratory health problems and exacerbating conditions like asthma and bronchitis. Additionally, it can damage crops and forests, contributing to ecological problems [1, 3, 4]. Thus, in the context of ozone level detection, it is essential to monitor ground-level ozone concentrations as part of assessing the quality of air and understanding the effects of ozone pollution on human health and the environment. By precisely detecting ozone levels, researchers can implement measures to mitigate ozone-related air pollution and its adverse effects.

Consequently, the accurate detection and monitoring of ozone levels has become a topic of paramount importance in environmental research and atmospheric science [1-8]. With the advent of machine learning techniques, researchers have gained powerful tools to analyze complex datasets and extract valuable insights from environmental data [1, 2, 4, 5, 8, 9]. However, to the best of the authors' knowledge, the gap relies in this research whereby feature selection has yet to be implemented in ozone level detection. Hence, it motivates this research to investigate the suitable feature selection method that can produce better classification results of ozone level detection.

Feature selection is a critical step in the machine learning pipeline which aims to identify the most relevant and informative features that contribute significantly to the target prediction task [10-12]. By eliminating irrelevant or redundant features, feature selection enhances model performance, reduces computational costs, and facilitates a deeper understanding of the underlying data [10-12]. Among the machine learning algorithms commonly employed in environmental studies, the kNN algorithm has gained popularity for its simplicity and effectiveness in classification tasks [11-13]. The kNN algorithm determines the class membership of a data point by considering the class labels of its k nearest neighbours [11-13].

Meanwhile, the Whale Optimization Algorithm (WOA) is a metaheuristic optimization technique inspired by the social behaviour of humpback whales [14]. Originally proposed for solving optimization problems, the WOA algorithm has demonstrated its capability to efficiently explore complex search spaces and find optimal solutions [15-18]. It draws inspiration from the foraging characteristic of humpback whales to solve optimization problems across various domains and is capable to handle various datasets including high-dimensional ones [19]. Not only that, a study presented a kNN-wrapped feature selection technique employing the WOA for classification purposes using Tournament and Roulette Wheel yielded good results while benchmarked against Particle Swarm Optimization (PSO), Ant Lion Optimizer (ALO) and Genetic Algorithm (GA) [20]. The outcomes of 98.2% accuracy in the Leukemia dataset highlighted the superiority of WOA in enhancing classification accuracy by efficiently identifying optimal feature subsets. In addition, a modified WOA (mWOA) was proposed for filter-based feature selection with inversed control parameter values [21]. The results of 90.75% average accuracy for the high-dimensional GLI_85 dataset proved that mWOA could perform better in terms of high precision rate and lower elapsed time as compared to the original WOA. The inversed control parameter technique also inspired another extension of work for wrapper-based mWOA and modified Grey Wolf Optimizer as well [22], whereby mWOA continued to show its ability in obtaining fewer number of selected features, higher accuracy by kNN, and shorter processing time. In conclusion, mWOA has shown its ability to process a variety of datasets, making it a promising candidate to detect ozone levels precisely and efficiently.

Therefore, this study aims to harness the potent capabilities of mWOA with the kNN classifier for refined feature selection and enhanced performance in the context of ozone level detection. By comparing the classification accuracy results obtained from the ozone level dataset without feature selection (Without FS), with features selected by the original WOA, and with features selected by mWOA, the proposed mWOA approach has achieved more accurate and efficient ozone level detection. The significance of this research resides in its comprehensive exploration and optimization of feature selection methodologies within the realm of ozone level detection leveraging mWOA. This study not only expands the horizons of mWOA beyond its prior applications in filter-based and wrapper-based feature selection but also establishes its clear superiority over the original WOA. Remarkably, the investigation reveals that the wrapper-based mWOA surpasses the performance of its filter-based counterpart, demonstrating higher accuracy and more refined feature selection. Furthermore, the research presents a nuanced consideration of the trade-off between time efficiency and accuracy by comparing filter-based and wrapper-based methodologies. In short, this study contributes to the advancement of ozone level detection by introducing and optimizing mWOA-driven feature selection methods.

The paper is organized into five sections: Section 1 provides an overview of ozone level detection and discusses previous related research and feature selection. Section 2 presents the related works regarding ozone level detection using machine learning approaches. Section 3 outlines the Whale Optimization Algorithm (WOA) along with its modified version (mWOA) for feature selection and Experimental Setup. Section 4 compares and discusses the experimental results of five techniques: dataset without feature selection, filter-based WOA, filter-based mWOA, wrapper-based WOA, and wrapper-based mWOA. Lastly,

Section 5 provides the conclusions drawn from the study's outcomes and proposes potential avenues for future research.

II. RELATED WORKS

Based on the literature, a study investigated the prediction of ground-level ozone to mitigate the health impacts of air pollution [9]. Multiple algorithms, including Support Vector Machines, Extreme Gradient Boosting (XGB), K-Nearest Neighbors (kNN), Hist Gradient Boosting Machine, and Deep Neural Networks, were employed to differentiate ozone and non-ozone days. It was identified that the XGB algorithm was the most suitable, achieving a 95% accuracy in detecting ozone layer concentrations. This work contributed to proactive health protection by forecasting ozone levels, endorsing XGB for precise ozone detection, and suggesting future exploration of broader datasets and attributes for enhanced applicability.

Besides, another study aimed to classify ground ozone levels as polluted or non-polluted based on big data by using machine learning models [8]. The dataset, sourced from the UCI Website contained atmospheric factors for Brazoria areas, underwent preprocessing, standardization, and division into training and testing sets. Employing Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and AdaBoost, the study evaluated model accuracy, with SVM achieving the highest test score of 94.9%. The research provided an essential forecasting approach for ground ozone pollution, showcasing the effectiveness of SVM among different models. Although challenges like missing values and small instances were noted, the study presented insights into model training and the potential utility of Neural Network and Deep Learning techniques for future predictions.

Moreover, a study was carried out focusing on ozone concentration prediction using machine learning models, taking into account the adverse effects of elevated ozone pollution on both human health and ecological systems [1]. Real data were utilized to assess prediction accuracy. While initial attempts using only meteorological variables did not manage to capture ozone trends, incorporating both meteorological and air pollutant data as input significantly improved model performance. The study highlighted results with 300% and 200% improvement in Root Mean Square Error compared to previous experiments. The best-performing model, specifically Support Vector Regression (SVR), demonstrated its effectiveness in ozone prediction, potentially benefiting air quality monitoring and management. The research suggested future investigations into graph models and complex spatiotemporal data for enhanced ozone prediction.

Furthermore, another optimization-related research was done by implementing hourly ozone simulation [5]. The study aimed to enhance surface ozone modelling in complex terrains, with a focus on Chongqing, a city plagued by ozone pollution and cloud cover in China. The proposed approach integrated ground-level information using ResNet (Residual Network), addressing challenges like vanishing gradients. A novel data structure treated ozone distribution as a non-still image, enhancing model performance. Cross-validation showed the robustness and feasibility of the approach, providing encouraging results for similar regions with cost-effectiveness and straightforward implementation.

Similarly, another regional research aimed to model ozone concentration's impact on air quality and public health during the winter season in Metropolitan Lima, Peru [2]. Exploratory, correlational, and predictive analyses were conducted using data from monitoring stations. LR and SVR were identified as effective predictive models. The study highlighted strong positive relationships between temperature and ozone, wind speed and ozone. Besides, ozone and relative humidity have a strong negative connection. This indicated the influence of anthropogenic activities and meteorological factors on urban air quality. The research provided a foundation for interventions in vulnerable areas, offering the potential for future ozone modelling applications. While further investigations and data are needed to enhance model accuracy and incorporate more variables, the study's findings hold relevance for environmental management and decision-making. It suggested avenues for future studies, including evaluating additional contaminants, employing varying coefficient models, and analyzing trends in meteorological variables and ozone concentration.

Notably, another work of ozone level prediction was done by researchers using a metaheuristic algorithm, Harris Hawk Optimization [4]. The objective of the research was to create a forecast model for ozone concentration employing Harris Hawk Optimization with SVR (HHO-SVR). Data from 14 JABODETABEK including ozone and meteorological factors, were collected. Recursive feature elimination with SVR (RFE-SVR) selected predictor variables, while the HHO-SVR model optimized SVR parameters

through Harris hawk optimization. Evaluation metrics confirmed the significance of lag variables, humidity, ultraviolet index, and air temperature, as predictors. Despite longer optimization times, the HHO-SVR model accurately predicted ozone concentration for 7 of 14 sub-districts, notably Ciputat and South Bekasi. The study recommended exploring imputation methods, precursor concentration data, and alternative metaheuristic algorithms for potential enhancements.

III. MATERIAL AND METHOD

The methodology section of this study comprises three subsections: (1) Whale Optimization Algorithm (WOA), (2) Modified Whale Optimization Algorithm for Feature Selection, and (3) Experimental Setup. In this section, a detailed overview of the two optimization algorithms, WOA and mWOA is presented. Additionally, the experimental setup used to evaluate the performance of WOA and mWOA in selecting features for ozone level detection has been described.

1. WHALE OPTIMIZATION ALGORITHM

The Whale Optimization Algorithm (WOA) was introduced by Mirjalili and Lewis in 2016 [14]. It stands as a metaheuristic approach inspired by a species called humpback whale, which has been acclaimed for its prowess in optimizing diverse domains including engineering, transportation, and medical diagnosis [15-28]. The algorithm's impressive performance can be attributed to its distinctive search mechanism, mimicking the bubble-net hunting behavior [23, 29], as illustrated in Figure 1.

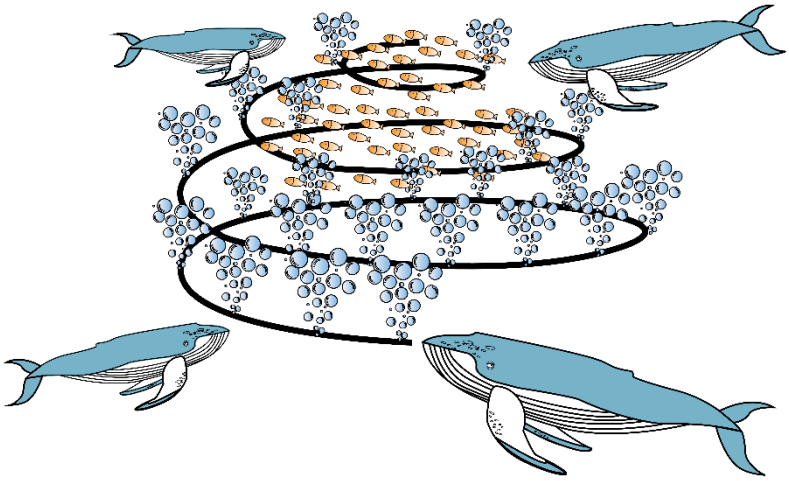

FIGURE 1. Illustration depicting the bubble-net hunting behavior.

A. SURROUNDING THE PREY

During this phase, the encirclement of the prey by the whales commences. The present optimal solution of the prey is assumed to be unknown, hence the best solution is initially considered as the prey's position. As the best position becomes evident, the whales progressively move towards this newfound optimum over successive iterations. This progression is mathematically captured by Equation (1) and Equation (2), where 't' signifies the as-is iteration, '$\vec{X}$' represents the position vector, '$\overrightarrow{X_{op}}$' corresponds to the position vector of the current best position, while '$||$' and '$\cdot$' denote the absolute value and dot product operator, respectively. With each iteration, the value of '$\overrightarrow{X*}$' undergoes updates whenever an improved position is discovered. '$\vec{D}$' represents the separation of the current position of the whale from the optimal position. The coefficient vectors '$\vec{A}$' and '$\vec{C}$' are formulated in Equation (3) and Equation (4), where '$\vec{r}$' signifies a random vector within

the range of 0 to 1, and '$\vec{a}$' denotes a control parameter which formulates values from 2 to 0 with equal intervals, following Equation (5), with ' $MaxIter$ ' representing the maximum number of iterations.

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X_{op}}(t) - \vec{X}(t)| \tag{1}$$

$$\vec{X}(t+1) = \overrightarrow{X_{op}}(t) - \vec{A} \cdot \vec{D} \tag{2}$$

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a} \tag{3}$$

$$\vec{C} = 2 \cdot \vec{r} \tag{4}$$

$$a = 2 - t\frac{2}{MaxIter} \tag{5}$$

B. BUBBLE-NET HUNTING

Bubble-net hunting is also known as the exploitation stage. During this stage, the whales engage in bubble-net attacks through two distinct maneuvers: the contraction of the encirclement and traversal spirally. The former is executed according to Equation (2), while the latter is effectuated utilizing Equation (6). Here, ' b ' serves as a constant that forms the spiral logarithmically whereas ' l ' represents a randomly generated value within the range of -1 to 1. ' $\overrightarrow{D'}$ ' is calculated as ' $|\overrightarrow{X_{op}}(t) - \vec{X}(t)|$ ', signifying the distance from the ith whale to the solution. As the whales execute these maneuvers concurrently, the probability ' p ' is introduced in Equation (7) to determine the chosen maneuver. If ' p ' is less than 0.5, the whales contract the circle; else, they embark on the spiral-shaped trajectory.

$$\vec{X}(t+1) = \overrightarrow{D'} \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X_{op}}(t) \tag{6}$$

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X_{op}}(t) - \vec{A} \cdot \vec{D} & if\ p < 0.5 \\ \overrightarrow{D'} \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X_{op}}(t) & if\ p \geq 0.5 \end{cases} \tag{7}$$

C. SEARCHING FOR PREY

Searching for prey is also known as the exploration stage. During this phase, whales engage in a search for potential positions, guided by randomness concerning their mutual positions. This distinct phase diverges from exploitation since a whale's position is transformed by referencing a randomly chosen counterpart, rather than relying on the best-found position. This phenomenon is encapsulated within Equation (8) and Equation (9), where ' $\overrightarrow{X_{rand}}$ ' signifies the position of a randomly picked whale from the population.

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X_{rand}} - \vec{X}| \tag{8}$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \cdot \vec{D} \tag{9}$$

In conclusion, WOA stands as a potent nature-inspired metaheuristic that harnesses the hunting patterns of humpback whales. By integrating both exploration and exploitation strategies through position updates based on optimal solutions and random selections, WOA emerges as an effective optimization tool applicable to diverse problem domains.

2. MODIFIED WHALE OPTIMIZATION ALGORITHM FOR FEATURE SELECTION

The introduction of the modified control parameter 'a' in WOA brings about a notable alteration in the optimization process. This parameter governs the balance between exploration and exploitation as the

algorithm progresses through iterations. By transitioning from the previous formula in Equation (5), to the updated form in Equation (10) [21], a more refined control mechanism emerges.

$$a = \frac{2\mathrm{t} + 2}{MaxIter} \tag{10}$$

Specifically, this modification enhances how whales explore the search area and adjust their positions. As iterations advance, parameter ' a ' contributes to a more balanced exploration and exploitation phases. The new formula ensures a linear increase in ' a ' from 0 to 2, facilitating a change from lesser search space to more search space at the beginning of the algorithm [21]. Conversely, at the end of the algorithm, the chance of position shifting is changed from high to low. This transition allows more thorough exploration in the beginning, followed by a more focused exploitation phase at the end.

Building upon the precursor work, this paper extends the application of mWOA. In the prior research, mWOA was used for filter-based feature selection in high-dimensional cancer datasets [21] and wrapper-based feature selection in general datasets [22]. This study further explores the utilization of mWOA in ozone level detection by comparing both feature selection techniques. This novel extension marks a significant step forward, showcasing the versatility and adaptability of mWOA across diverse domains.

The proposed mWOA is developed for feature selection in both filter-based and wrapper-based techniques. In the wrapper-based technique, the fitness value is computed through k-Nearest Neighbors (kNN), while the filter-based method derives the fitness value from Equation (11). This equation was adapted from another filter-based WOA used in cancerous cell detection [26]. Based on Equation (11), variables ' x ' and ' y ' represent the random samples for Label 0 (Normal) and Label 1 (Ozone), whereas ' $MeanNormal$ ' and ' $MeanOzone$ ' indicate the average value of the samples for each label. The fitness value symbolized the distance between ' x ' and ' y '. The higher the value generated, the fitter the solution is.

$$fitness = \sqrt{(x - MeanNormal)^2 + (y - MeanOzone)^2} \tag{11}$$

The difference between the proposed filter-based and wrapper-based techniques extends beyond fitness value formation; it encompasses the number of selected features as well. In the wrapper-based technique, the selection relies on fitness value, with features chosen if their fitness value falls outside a specific threshold, which is 0.5 in this case. Conversely, in the filter-based approach, all features are sorted based on the generated fitness value, and only the (fitter) upper half is chosen for inclusion. This filter-based approach was inspired by a precursor of this work, where eliminating half of the HDD features yields optimal classification accuracy [21]. In short, mWOA was implemented in wrapper-based and filter-based techniques which highlighted how versatile mWOA is for improving feature selection in different situations. Figure 2 illustrates the pseudocode detailing mWOA-driven feature selection where n denotes number of features, t represents iteration counter, and $maxIter$ indicates maximum number of iterations = 100. For the experiment, this procedure is applicable to filter-based and wrapper-based feature selection methods.

Step 1: Initialize the whale population X_i $(i = 1, 2, \ldots, n)$.
Step 2: Compute each search agent's fitness: Equation (11) for filter-based; kNN for wrapper-based.
Step 3: Assign X_{op} = the best search agent
Step 4: Finding the optimal solution throughout each iteration

```
While (t < maxIter)
  for each search agent
    Update a, A, C, with Equations (10), (3), (4), set l = [−1,1] and p = [0,1] respectively
    if (p < 0.5)
      if (|A| < 1)
        Update the current search agent's position by Equation (2)
      else
        Choose a search agent by random (X_rand)
        Update the current search agent's position by Equation (9)
      end if
    else
```

Update the current search agent's position by Equation (6)
end if
end for
Fix the search agent's position if it outstrips the search space
Compute each search agent's fitness: Equation (11) for filter-based; kNN for wrapper-based.
Update X_{op} if a fitter solution is found
$t = t + 1$
end while
Step 5: Feature selection
For filter-based, select 50% features based on fitness value.
For wrapper-based, select features with fitness value > 0.5 threshold.

FIGURE 2. Pseudocode of mWOA-driven feature selection.

3. EXPERIMENTAL SETUP

The software and hardware configuration for the experiment is as follows.

- Operating System: The experiments were conducted on a computer running Windows 11 version 22H2, a 64-bit operating system.
- Processor: The computer was equipped with an Intel Core i7-10750H CPU with a base clock speed of 2.60 GHz.
- RAM: A total of 32.0 GB RAM was installed, with 31.8 GB being usable during the experiments.
- GPU: The experimental setup included an NVIDIA GeForce RTX 3060 Laptop GPU for accelerated computations.
- MATLAB: The experiments utilized MATLAB as the primary programming and analysis platform. The version employed was MATLAB R2023a.

The Ozone Level Detection dataset is retrieved from the UCI machine learning repository [30]. The attributes of the dataset are in numerical values. The dataset has the characteristic of multivariate where it is described by more than one feature, some of the features include temperature, humidity, wind direction, geopotential height, and sea level pressure. Besides, the dataset has characteristics of time series, where data points are sampled eight-hourly in sequential order. The dataset has 2535 observations, 72 features, and a binary class whereby Label 0 indicates normal and Label 1 represents harmful ozone. However, only 1847 of the observations are without missing values. Therefore, this study performed feature selection and classification based on 1847 observations for ozone level detection.

The experimental procedure involved evaluating the effectiveness of the mWOA for feature selection in the context of ozone level detection, comparing it with the original WOA using both filter-based and wrapper-based approaches. There were five experiment techniques used as follows.

- Without FS: The complete dataset was used for classification using the kNN model, without any feature selection.
- Filter-based WOA: The original WOA was applied with a filter-based approach to select 50% of the most relevant features based on the fitness value calculated in Equation (11).
- Filter-based mWOA: The proposed mWOA was employed with a filter-based approach to select 50% of the most relevant features based on the fitness value calculated by Equation (11).
- Wrapper-based WOA: The original WOA was applied with a wrapper-based approach, where the algorithm selected suitable amounts of features based on the fitness value calculated by the kNN-trained model using different feature subsets.
- Wrapper-based mWOA: The proposed mWOA was employed with a wrapper-based approach, where the algorithm selected suitable amounts of features based on the fitness value calculated by the kNN-trained model using different feature subsets.

The K-Nearest Neighbors (kNN) classification model was employed for predicting ozone levels based on the selected features. The parameter settings of kNN are as follows: the 'HoldOut' validation technique was employed with a value of k set to 5; the dataset was partitioned into training and validation sets using a ratio

of 0.2, ensuring that 20% of the data was reserved for validation purposes while 80% of the data was used in training. The experiments were repeated 20 times to ensure reliable and consistent results. For each experiment run under the five techniques, the following evaluation metrics were recorded:

- Processing Time: The time in unit seconds taken to execute the feature selection process and subsequent kNN classification was measured for each experiment to assess computational efficiency.
- Number of Features Selected: The number of chosen features for each scenario was documented to grasp the influence of feature selection on the model's performance.
- Classification Accuracy: The accuracy percentage of the kNN model in predicting ozone levels using the selected features was evaluated for each run. This performance metric allowed for a comparison of the model's effectiveness under different feature selection techniques. The accuracy of each trial was computed utilizing Equation (12) wherein TP (True Positive) and TN (True Negative) correspond to the count of accurately classified positive and negative instances, while FP (False Positive) and FN (False Negative) represent the count of incorrect classified positive instances and negative instances, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

Upon completing the 20 experiment runs for each technique, the minimum, maximum, and mean of the mentioned metrics (number of features selected, processing time, and classification accuracy) were calculated. This comprehensive analysis provided a clear understanding of the impact of using mWOA for both filter-based and wrapper-based feature selection on ozone level detection performance. The statistical results will be used as a foundation for deriving significant conclusions and understanding the strengths and benefits of the proposed mWOA approach over alternative methods, which will be discussed in the forthcoming Results and Discussion section.

IV. RESULT AND DISCUSSION

This section presents a comprehensive evaluation of the proposed techniques for ozone level detection. This section is divided into subsections focusing on three key evaluation criteria: processing time, number of features selected, and classification accuracy. The obtained statistical metrics, including minimum, maximum, and mean, are thoroughly analyzed to assess the performance of the WOA and mWOA.

1. PROCESSING TIME

The comparison of processing time for various feature selection techniques in Table 1 provides valuable insights into the processing time efficiency of different feature selection techniques in the context of ozone level detection.

Table 1. Comparison of processing time (s).

Techniques	Minimum	Maximum	Mean
Without FS	**# 0.0078788**	**# 0.0117073**	**# 0.0089632**
Filter-based WOA	0.2268366	0.3274791	0.2441505
Filter-based mWOA	*** 0.2249756**	*** 0.2691878**	*** 0.2397550**
Wrapper-based WOA	5.1148457	7.0697780	6.0286101
Wrapper-based mWOA	*** 4.1439817**	*** 5.3859644**	*** 4.8680684**

Note: (*) = mWOA outperformed WOA; (#) = best results

The processing time includes the entire process of feature selection steps (if any) and classification steps. Four techniques were compared: Without Feature Selection (Without FS), Filter-based Whale Optimization Algorithm (WOA), Filter-based modified Whale Optimization Algorithm (mWOA), Wrapper-based WOA, and Wrapper-based mWOA. The evaluation metrics are statistically measured based on the minimum, maximum, and mean values of 20 experiments.

Observing the processing time values based on Table 1, it is evident that the Without FS approach achieved the shortest processing time, with a range of 0.0078788 to 0.0117073 seconds. This is expected, as Without FS directly utilizes the entire dataset with classification without involving any feature selection process, resulting in the fastest execution with a mean value of 0.0089632 seconds. However, it is important to note that the Without FS approach does not consider feature relevance, potentially leading to suboptimal model performance due to irrelevant or redundant features.

Moving to the filter-based and wrapper-based approaches, it can be observed that both WOA and mWOA introduce additional processing time as compared to Without FS due to involving the feature selection step. Among the filter-based approaches, the mWOA technique demonstrates slight improvement over the original WOA, with minimum, maximum, and mean processing times of 0.2249756, 0.2691878, and 0.2397550 seconds, respectively. On average, filter-based mWOA outperformed WOA by 0.0043955 seconds. Similarly, in the wrapper-based category, mWOA proves to be more efficient than WOA, with processing times ranging from 4.1439817 to 5.3859644 seconds and a mean of 4.8680684 seconds. On average, wrapper-based mWOA outperformed WOA by 1.1605417 seconds.

The results indicated by (*) in Table 1 highlight instances where mWOA outperformed WOA in terms of processing time. These improvements can be attributed to the modified control parameter of mWOA, which allows for a more precise exploration of the search area during feature selection. The ability of mWOA to achieve comparable or better results in less time than WOA makes it a more efficient and promising choice for ozone level detection tasks. Overall, the discussion of processing time reveals that the mWOA-based feature selection approach showcases computational efficiency while maintaining or improving the quality of selected features. This efficiency is crucial, especially when dealing with large datasets and real-time applications, where reducing processing time is of utmost importance.

2. NUMBER OF SELECTED FEATURES

The results presented in Table 2 offer valuable insights into the number of selected features by different feature selection techniques for ozone level detection. For the Without FS approach, all 72 features from the original dataset were used without any selection. While this approach provides a complete set of features, it may include insignificant ones, potentially leading to reduced model performance.

Table 2. Comparison of the number of selected features.

Techniques	Minimum	Maximum	Mean
Without FS	72	72	72
Filter-based WOA	36	36	36
Filter-based mWOA	36	36	36
Wrapper-based WOA	4	26	12.1
Wrapper-based mWOA	**#* 2**	**#* 20**	**#* 7.85**

Note: (*) = mWOA outperformed WOA; (#) = best results

In the filter-based approaches, both WOA and mWOA selected 50% (36 features) of the total features from the dataset. The selection process involved sorting all 72 features based on their fitness value, with fitter features deemed more relevant. As a result, the top half of the features with the highest fitness values were chosen for inclusion in the model. Hence, the results of both filter-based techniques yielded the same number of selected features.

Conversely, the wrapper-based approaches exhibited more variability in the number of selected features. The wrapper-based WOA technique selected between 4 to 26 features, with a mean of 12.1, indicating a flexible feature subset based on the evaluation of the kNN model's performance. Notably, the wrapper-based mWOA outperformed WOA in terms of feature selection as indicated in Table 2. The wrapper-based mWOA technique selected between 2 to 20 features, with a mean of 7.85.

On average, the wrapper-based mWOA chose 4.25 fewer features than WOA. This difference is attributed to the modified control parameter in mWOA, which facilitated a more precise exploration of the search space during feature selection. As a result, mWOA identified a smaller yet more informative set of features

compared to WOA in the wrapper-based approach. In conclusion, the discussion of the number of selected features confirms the effectiveness of mWOA in achieving more compact and informative feature subsets.

3. CLASSIFICATION ACCURACY

The comparison of classification accuracy for various feature selection techniques is shown in Table 3.

Table 3. Comparison of classification accuracy.

Techniques	Minimum	Maximum	Mean
Without FS	92.95	92.95	92.95
Filter-based WOA	92.95	92.95	92.95
Filter-based mWOA	*** 93.50**	*** 93.77**	*** 93.75**
Wrapper-based WOA	**# 93.77**	94.85	94.31
Wrapper-based mWOA	93.23	**#* 95.40**	**#* 94.49**

Note: (*) = mWOA outperformed WOA; (#) = best results

The Without FS and the filter-based WOA approaches both obtained the same classification accuracy of 92.95%. This suggests that using the entire dataset without feature selection or selecting 50% of the features based on the fitness values by WOA yielded similar predictive performance.

In contrast, the mWOA approach in the filter-based category outperformed WOA in terms of classification accuracy, as indicated by (*). The minimum, maximum, and mean accuracy values for filter-based mWOA were 93.50%, 93.77%, and 93.75%, respectively. On average, filter-based mWOA performed 0.80% better than WOA. This highlights the effectiveness of the modified control parameter in mWOA, which contributed to a slight improvement in the model's predictive power.

Regarding the wrapper-based approaches, the wrapper-based WOA achieved higher classification accuracy than the filter-based techniques, ranging from 93.77% to 94.85%, with a mean accuracy of 94.31%. Additionally, the wrapper-based mWOA demonstrated promising results, with accuracy values ranging from 93.23% to 95.40% and a mean accuracy of 94.49%. On average, wrapper-based mWOA performed 0.18% better than WOA. The performance of wrapper-based mWOA further emphasizes the significance of the modified control parameter, enabling mWOA to identify a more informative set of features, leading to improved model accuracy.

In conclusion, the discussion of classification accuracy reveals the advantages of using mWOA for feature selection in ozone level detection. The results underscore the superior predictive power of mWOA in both filter-based and wrapper-based approaches, as compared to the original WOA method. The statistical analysis validates mWOA's potential as an effective feature selection technique, contributing to enhanced classification accuracy and ultimately leading to more accurate ozone level prediction models.

V. CONCLUSION

This study presented a comprehensive investigation of ozone level detection through feature selection, employing the modified Whale Optimization Algorithm (mWOA). The research aimed to enhance the accuracy and efficiency of ozone level prediction models by identifying the most informative features from the dataset. Through extensive experiments and evaluations, the effectiveness of the mWOA-based feature selection approach was thoroughly analyzed and compared with the original Whale Optimization Algorithm (WOA).

Without feature selection, the processing time is the least. However, feeding all 72 features to the model does not help to improve the accuracy of ozone level detection. By introducing WOA and mWOA as feature selection techniques, their performances are tested in both filter-based and wrapper-based techniques. For filter-based techniques, although both mWOA and WOA selected the same number of features, their processing time and classification accuracy are not the same. Based on the filter-based results, mWOA outperformed WOA by 0.0043955 seconds less in terms of average processing time. Furthermore, filter-based mWOA exhibited an average classification accuracy that was 0.80% higher than that of filter-based WOA. As for wrapper-based techniques, the results demonstrated that mWOA outperformed WOA in terms of processing time, number of selected features, and classification accuracy. Comparing the wrapper-based

results, mWOA chose 4.25 fewer features than WOA on average. Besides, wrapper-based mWOA outperformed wrapper-based WOA by 1.1605417 seconds less in terms of average processing time. Moreover, wrapper-based mWOA exhibited an average classification accuracy that was 0.18% higher than that of wrapper-based WOA.

According to the superior results of mWOA in both filter-based and wrapper-based techniques, the modified control parameter of mWOA enabled more precise exploration of the search space during feature selection, leading to improved efficiency and performance. The statistical metrics revealed that mWOA effectively selected a more concise and informative subset of features in less time, contributing to enhanced model accuracy without compromising computational complexity. The reduced number of selected features further supported the computational efficiency and real-world applicability of the mWOA approach. Nevertheless, while mWOA adeptly retains its effectiveness as highlighted by the experimental outcomes, it encounters the challenge of upholding the stability of its search strategy during exploration and exploitation.

In conclusion, the findings of this study indicate that mWOA is a promising and efficient feature selection method for ozone level detection tasks. Its ability to improve classification accuracy, reduce processing time, and identify a concise set of informative features highlights its potential for practical applications in environmental research and atmospheric science. However, the limitation of mWOA is that the modification on the control parameter alone is insufficient to produce significant difference in the performance compared to original WOA. Future research may explore the application of mWOA in other environmental monitoring and prediction domains and extend the study to investigate its performance with different machine learning classifiers. Furthermore, addressing the enhancement of mWOA's stability in navigating exploration and exploitation phases stands as a prospective challenge, warranting investigation for further refinement in future research.

ACKNOWLEDGEMENT

Communication of this research is made possible through monetary assistance by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216. This research also was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through GPPS (vot Q290).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Pan, Q., Harrou, F., & Sun, Y., "A comparison of machine learning methods for ozone pollution prediction," *Journal of Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00748-x.
2. Carbo-Bustinza, N. *et al.*, "A machine learning approach to analyse ozone concentration in metropolitan area of Lima, Peru," *Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022, doi: 10.1038/s41598-022-26575-3.
3. Antipova, A., *Analysis of exposure to ambient air pollution: Case study of the link between environmental exposure and children's school performance in Memphis, TN*. Elsevier Inc., 2020. doi: 10.1016/b978-0-12-815822-7.00011-x.
4. Kurniawan, R., Setiawan, I. N., Caraka, R. E., & Nasution, B. I., "Using Harris hawk optimization towards support vector regression to ozone prediction," *Stochastic Environmental Research and Risk Assessment*, vol. 36, no. 2, pp. 429–449, 2022, doi: 10.1007/s00477-022-02178-2.
5. Zhu, S. *et al.*, "An Optimization Approach for Hourly Ozone Simulation: A Case Study in Chongqing, China," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 1871–1875, 2021, doi: 10.1109/LGRS.2020.3010416.
6. Kwan, S. C. *et al.*, "Health impacts from TRAPs and carbon emissions in the projected electric vehicle growth and energy generation mix scenarios in Malaysia," *Environmental Research*, vol. 216, no. P2, p. 114524, 2023, doi: 10.1016/j.envres.2022.114524.
7. Usmani, R. S. A., Saeed, A., Abdullahi, A. M., Pillai, T. R., Jhanjhi, N. Z., & Hashem, I. A. T., "Air pollution and its health impacts in Malaysia: a review," *Air Quality, Atmosphere and Health*, vol. 13, no. 9, pp. 1093–1118, 2020, doi: 10.1007/s11869-020-00867-x.
8. Mohammed, M. A., Lakhan, A., Zebari, D. A., Abdulkareem, K. H., Nedoma, J., Martinek, R., ... & Tiwari, P. (2023). Adaptive secure malware efficient machine learning algorithm for healthcare data. *CAAI Transactions on Intelligence Technology*.
9. Sarkar, A., Ray, S. S., Prasad, A., & Pradhan, C., "A Novel Detection Approach of Ground Level Ozone using Machine Learning Classifiers," *Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2021*, pp. 428–432, 2021, doi: 10.1109/I-SMAC52330.2021.9640852.
10. Aighuraibawi, A. H. B., Manickam, S., Abdullah, R., et al. (2023). Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm. *Comput. Syst. Sci. Eng.*, *47*(1), 553-574.

11. Ali, R. R., Mohamad, K. M., Mostafa, S. A., et al. (2023). A meta-heuristic method for reassemble bifragmented intertwined JPEG image files in digital forensic investigation. *IEEE Access*.
12. Madhusudhanan, B., Sumathi, P., Karpagam, N. S., Mahesh, A., & Suhi, P. A. P., "An hybrid metaheuristic approach for efficient feature selection," *Cluster Computing*, vol. 22, no. s6, pp. 14541–14549, 2019, doi: 10.1007/s10586-018-2337-2.
13. Taher, K. I., Abdulazeez, A. M., & Zebari, D. A. (2021). Data mining classification algorithms for analyzing soil data. *Asian Journal of Research in Computer Science*, *8*(2), 17-28.
14. Mirjalili, S. & Lewis, A., "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016, doi: 10.1016/j.advengsoft.2016.01.008.
15. Nadimi-Shahraki, M. H., Zamani, H., & Mirjalili, S., "Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study," *Computers in Biology and Medicine*, vol. 148, no. January, p. 105858, 2022, doi: 10.1016/j.compbiomed.2022.105858.
16. Sony, B., Chakravarti, A., & Reddy, M. M., "Traffic congestion detection using whale optimization algorithm and multi-support vector machine," *International Journal of Recent Technology and Engineering*, vol. 7, no. 6C2, pp. 589–593, 2019.
17. Brodzicki, A., Piekarski, M., & Jaworek-Korjakowska, J., "The whale optimization algorithm approach for deep neural networks," *Sensors*, vol. 21, no. 23, 2021, doi: 10.3390/s21238003.
18. Hussien, A. G., Oliva, D., Houssein, E. H., Juan, A. A., & Yu, X., "Binary whale optimization algorithm for dimensionality reduction," *Mathematics*, vol. 8, no. 10, pp. 1–24, 2020, doi: 10.3390/math8101821.
19. Yab, L. Y., Wahid, N., & Hamid, R. A., "A Meta-Analysis Survey on the Usage of Meta-Heuristic Algorithms for Feature Selection on High-dimensional Datasets," *IEEE Access*, vol. 10, no. November, pp. 1–1, 2022, doi: 10.1109/access.2022.3221194.
20. Mafarja, M. & Mirjalili, S., "Whale optimization approaches for wrapper feature selection," *Appl Soft Comput*, vol. 62, pp. 441–453, 2018, doi: 10.1016/j.asoc.2017.11.006.
21. Yab, L. Y., Wahid, N., & Hamid, R. A., *A Modified Whale Optimization Algorithm as Filter-Based Feature Selection for High Dimensional Datasets*, vol. 457 LNNS. Springer International Publishing, 2022. doi: 10.1007/978-3-031-00828-3_9.
22. Yab, L. Y., Wahid, N., & Hamid, R. A., "Inversed Control Parameter in Whale Optimization Algorithm and Grey Wolf Optimizer for Wrapper-Based Feature Selection: A Comparative Study," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 477–486, 2023, doi: 10.30630/joiv.7.2.1509.
23. Gharehchopogh, F. S. & Gholizadeh, H., "A comprehensive survey: Whale Optimization Algorithm and its applications," *Swarm Evol Comput*, vol. 48, no. November 2018, pp. 1–24, 2019, doi: 10.1016/j.swevo.2019.03.004.
24. Saidala, R. K., & Devarakonda, N., "Improved whale optimization algorithm case study: Clinical data of anaemic pregnant woman," *Advances in Intelligent Systems and Computing*, vol. 542, pp. 271–281, 2018, doi: 10.1007/978-981-10-3223-3_25.
25. Chakraborty, S., Saha, A. K., Nama, S., & Debnath, S., "COVID-19 X-ray image segmentation by modified whale optimization algorithm with population reduction," *Computers in Biology and Medicine*, vol. 139, no. October, p. 104984, 2021, doi: 10.1016/j.compbiomed.2021.104984.
26. Nematzadeh, H., Enayatifar, R., Mahmud, M., & Akbari, E., "Frequency based feature selection method using whale algorithm," *Genomics*, vol. 111, no. 6, pp. 1946–1955, 2019, doi: 10.1016/j.ygeno.2019.01.006.
27. Yuan, X., Miao, Z., Liu, Z., Yan, Z., & Zhou, F., "Multi-strategy ensemble whale optimization algorithm and its application to analog circuits intelligent fault diagnosis," *Applied Sciences (Switzerland)*, vol. 10, no. 11, 2020, doi: 10.3390/app10113667.
28. Tan, W. H., & Mohamad-Saleh, J., "A hybrid whale optimization algorithm based on equilibrium concept," *Alexandria Engineering Journal*, vol. 68, pp. 763–786, 2023, doi: 10.1016/j.aej.2022.12.019.
29. Rana, N., Latiff, M. S. A., Abdulhamid, S. M., & Chiroma, H., *Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments*, vol. 32, no. 20. Springer London, 2020. doi: 10.1007/s00521-020-04849-z.
30. Zhang, K., Fan, W., & Yuan, X., "Ozone Level Detection," UC Irvine Machine Learning Repository. Accessed: Jul. 20, 2023. Available: https://archive.ics.uci.edu/dataset/172/ozone+level+detection