Journal of Agricultural Sciences
(Tarim Bilimleri Dergisi)

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

JAS

A Meta-Heuristic Algorithm-Based Feature Selection Approach to Improve Prediction Success for *Salmonella* Occurrence in Agricultural Waters

Murat DEMIR[a], Murat CANAYAZ[b], Zeynal TOPALCENGIZ [c, d*]

[a]*Department of Software Engineering, Faculty of Engineering and Architecture, Muş Alparslan University, 49250 Muş, TÜRKİYE*

[b]*Department of Computer Engineering, Faculty of Engineering, Van Yüzüncü Yıl University, 65080 Van, TÜRKİYE*

[c]*Department of Food Science, Center for Food Safety, University of Arkansas System Division of Agriculture, Fayetteville, AR 72704, USA*

[d]*Department of Food Engineering, Faculty of Engineering and Architecture, Muş Alparslan University, 49250 Muş, TÜRKİYE*

ARTICLE INFO

Research Article
Corresponding Author: Zeynal TOPALCENGIZ, E-mail: zeynalt@uark.edu; zeynaltopalcengiz@gmail.com
Received: 24 May 2023 / Revised: 20 September 2023 / Accepted: 21 September 2023 / Online: 09 January 2024

Cite this article

Demir M, Canayaz M, Topalcengiz Z (2024). A Meta-Heuristic Algorithm-Based Feature Selection Approach to Improve Prediction Success for *Salmonella* Occurrence in Agricultural Waters. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 30(1):118-130. DOI: 10.15832/ankutbd.1302050

ABSTRACT

The presence of *Salmonella* in agricultural waters may be a source of produce contamination. Recently, the performances of various algorithms have been tested for the prediction of indicator bacteria population and pathogen occurrence in agricultural water sources. The purpose of this study was to evaluate the performance of meta-heuristic optimization algorithms for feature selection to increase the *Salmonella* occurrence prediction success of commonly used algorithms in agricultural waters. Previously collected datasets from six agricultural ponds in Central Florida included the population of indicator microorganisms, physicochemical water attributes, and weather station measurements. *Salmonella* presence was also reported with PCR-confirmed method in data set. Features were selected by using binary meta-heuristic optimization methods including differential evolution optimization (DEO), grey wolf optimization (GWO), Harris hawks optimization (HHO) and particle swarm optimization (PSO). Each meta-heuristic method was run 100 times for the extraction of features before classification analysis. Selected features after optimization were used in the K-nearest neighbor algorithm (kNN), support vector machine (SVM) and decision tree (DT) classification methods. Microbiological indicators were ranked as the first or second features by all optimization algorithms. Generic *Escherichia coli* was selected as the first feature 81 and 91 times out of 100 using GWO and DEO, respectively. The meta-heuristic optimization algorithms for the feature selection process followed by machine learning classification methods yielded a prediction accuracy between 93.57 and 95.55%. Meta-heuristic optimization algorithms had a positive effect on improving *Salmonella* prediction success in agricultural waters despite spatio-temporal variations. This study indicates that the development of computer-based tools with improved meta-heuristic optimization algorithms can help growers to assess risk of *Salmonella* occurrence in specific agricultural water sources with the increased prediction success.

Keywords: Optimization, Support Vector Machine, kNN, Decision tree, Water quality

1. Introduction

Agricultural waters can be the main source of microbiological contamination in produce fields (FDA 2015). Pathogens such as *Salmonella* and shiga toxin-producing *Escherichia coli* can survive at various temperatures in agricultural surface waters for prolonged periods of time (Topalcengiz & Danyluk 2019; Topalcengiz et al. 2019). Agricultural water sources have been implicated as the possible source of *Salmonella* contamination during produce related outbreaks (CDC 2007; Greene et al. 2008). Microbiological indicators including streptococci, enterococci, and total coliforms can be used to monitor the water quality (Steele et al. 2005). The measurement of a generic *Escherichia coli* population is commonly required or recommended to assess the risk of contamination from agricultural water sources around the world (Ashbolt 2001; FDA 2015). However, weather conditions and environmental factors may cause dramatic changes in agricultural surface water quality that may increase the risk of produce contamination.

Computer-based tools have been recently used to analyze the microbiological quality of agricultural water sources with various algorithms (Abimbola et al. 2020; Weller et al. 2020; Buyrukoğlu 2021; Buyrukoğlu et al. 2021). Artificial neural networks (ANN), K-Nearest neighbor algorithm (kNN), support vector machine (SVM), decision tree, random forest and AdaBoost can be listed as the most preferred algorithms to predict the presence of *Salmonella* based on measured environmental factors, the population of microbiological indicators, and the physicochemical attributes of agricultural waters (Polat et al. 2020; Weller et al. 2020; Buyrukoğlu 2021). In published studies, the performance of computer-based tools is mainly evaluated with the value of accuracy with or without feature selection.

Feature selection is considered a critical step towards improving the prediction success of machine learning tools by eliminating inappropriate, irrelevant, or unnecessary features (Agrawal et al. 2021). Meta-heuristic methods provide effective and acceptable solution methods for future selection-based optimization. Solutions are candidate values that can be a set of desired outputs for each method to get closer to better results depending on the structure of the algorithms. Meta-heuristic algorithms consist of two main components: intensification and diversification (Blum & Roli 2003). Intensification focuses on producing a solution in a local area with the best available solution. Diversification means creating a variety of solutions to explore the search space on a global scale. The combined selection of the best solutions ensures that the solutions converge towards the optimum (Yang 2011). Diversification also prevents solutions from being localized and increases the diversity of solutions to avoid stagnation in local optima or flat areas. Each algorithm uses different methods to achieve a balance between concentration and diversification.

The presence and concentration of pathogens in agricultural water have been predicted with artificial intelligence and machine learning tools with various classification techniques. In general, the success of classification techniques is evaluated with or without feature selection based on a researcher's preferences, experiences, data availability, and algorithm popularity. Feature selection can be performed with statistical and computer-based tools. In a recent survey and review, grey wolf optimization (GWO), Harris hawks optimization (HHO), differential evolution optimization (DEO), and particle swarm optimization (PSO) have been listed as the most studied conventional metaheuristic algorithms used on data sets produced in various fields (Akinola et al. 2022; Dokeroglu et al. 2022). In this study, the effectiveness of above-mentioned meta-heuristic optimization algorithms was evaluated for feature selection to improve the *Salmonella* occurrence prediction performance of commonly used algorithms in agricultural waters.

2. Material and Methods

2.1. Data set

A previously acquired data set for six agricultural ponds from Central Florida was obtained by Topalcengiz et al. (2017). The data set included the population of indicator microorganisms (total coliform, generic *Escherichia coli*, and enterococci), physicochemical attributes of water samples (air and water temperature, pH, oxidative reduction potential, conductivity, and turbidity), and weather station measurements as rain and solar radiation for 24 h before sampling, average solar radiation, 60 cm air temperature, relative humidity, ten-meter wind speed, wind direction, and 60 cm soil temperature in total of 540 samples (90 from each pond) for two growing seasons. In addition, the presence of *Salmonella* in water samples was confirmed through PCR after enrichment. In this study, adjustments were made on the data set at hand. In this context, a data set of 540 values * 17 features was obtained by combining all data from the six ponds. The class label for this dataset was determined as 1 for the presence of the *Salmonella* pathogen and 0 for its absence. Input and output values were normalized in the range of 0-1. Equation 1 was used for normalization.

$$y = (x_i - x_{min})/(x_{max} - x_{min}) \quad (1)$$

Where; y is the normalized value of x_i. The x_{max} and the x_{min} are the maximum and minimum value of x_i, respectively.

2.2. Methods

Two process steps were applied. First, the feature selection was staged. After normalization, the data set was subjected to feature selection through four different meta-heuristic optimization algorithms. Metaheuristic algorithms including differential evolution optimization (DEO), grey wolf optimization (GWO), Harris hawks optimization (HHO) and particle swarm optimization (PSO) were assessed based on the nearest neighborhood algorithm (kNN) with 5 neighborhoods as a fitness function. The rate of error obtained from the kNN was checked each time after each run. When the error rate was lower than the previous value, the features providing this value were taken as the best values. The population size was standardized as 20 with 100 iterations for comparison of all tested meta-heuristic algorithms.

In the classification phase, the dataset was first segmented by using cross validation with k value of 5. One of these parts was used as test data for the part classification algorithm. The cross-validation method was used to confirm the reliability and accuracy of the results in the studies. The data sets obtained with k-fold are classified by support vector machines (SVM), kNN and decision tree algorithms based on successful classification as described in previous studies conducted on the same data set, respectively (Polat et al. 2020; Buyrukoğlu 2021; Buyrukoğlu et al. 2021). During the application, fitcsvm, fitcknn, fitctree functions in the Matlab program were used for classification. Default values were used for fitcsvm. NumNeighbors:5, Distance:minkowski parameters were used for fitcknn. Finally, MaxNumSplits: 7 value was applied for fitctree.

2.3. Feature selection

The multidimensionality of the data is considered as a challenge for classification techniques as well as for all data mining the methods. A reduction in the number of classified dimensions reduces computational demands and data collection requests with

increase in reliability of baseline results and data quality. In this study, binary versions of DEO, GWO, HHO, and PSO meta-heuristic methods were selected for feature selection to increase the accuracy success of classifiers. These meta-heuristic methods were determined based on previous successful applications by the authors (Canayaz 2021) and frequent use in the literature (Akinola et al. 2022; Dokeroglu et al. 2022). In this respect, it is also possible to evaluate our study as an ablation study.

2.3.1. Binary differential evolution optimization

The differential evolution (DEO) algorithm is a widely used as population-based stochastic direct search method for solving continuous-time optimization problems (Storn & Price 1997; Price et al. 2005; Das & Suganthan 2011). It uses real number coding and involves three basic operations: mutation, crossover, and selection. The initial population is randomly generated and covers the entire search space. While the traditional DEO algorithm is effective at solving continuous-time problems, it is unable to handle discrete problems and does not consider global or neighboring individual solution information. In contrast, the binary DEO incorporates information from neighboring solutions during the crossover phase to improve its performance on discrete problems (Liang et al. 2017). Binary DEO operates differently from the traditional DEO algorithm in the population initialization, mutation, and crossover phases.

Binary DEO creates the initial population with the formula in Equation 2 (Liang et al. 2017):

$$\begin{cases} 1, & rand(0,1) < 0.05 \\ 0, & otherwise \end{cases} \quad (2)$$

The mutation operator was performed with the formula in Equation 3:

$$V_{i,G}^{j} = \begin{cases} x_{p1,G}^{j} \mid x_{p2,G}^{j} & x_{p1,G}^{j} = x_{p2,G}^{j} \\ x_{i,G}^{j} & \text{otherwise} \end{cases} \quad (3)$$

For the jth candidate node, if individuals $X_{p1,G}$, $X_{p2,G}$ have the same choice, the mutant individuals yields $x_{p1,G}^{j}$ or $x_{p2,G}^{j}$, otherwise it directly derives form $X_{i,G}$.

Crossover Operator was performed by the formula in Equation 4:

$$V_{i,G}^{j} = \begin{cases} v_{nbest,G}^{j} & rand_j[0,1) \leq CR \text{ or } j = rand(i) \\ v_{i,G}^{j} & \text{otherwise} \end{cases} \quad (4)$$

The crossover ratio CR was chosen by the designer in the range [0,1). Crossover ensures that $U_{i,G}$ has at least one value from the best neighbour. Neighborhood radius r depends on population size and complexity of the problem.

2.3.2. Binary grey wolf optimization

The binary version of grey wolf optimization (GWO) is an optimization algorithm inspired by the hunting and social behavior of grey wolves (Mirjalili et al. 2014). It involves a group of 5-12 wolves, divided into four categories: alpha, beta, delta, and omega. The alpha wolf is the leader and makes decisions concerning hunting, sleep times, and sleeping locations. The beta wolf assists the alpha wolf, while the delta wolf follows the alpha and beta wolves and only dominates the omega wolf, the lowest ranking member of the group. In this study, the binary version of GWO was used for feature selection, with the kNN error rate serving as the fitness function (Emary et al. 2016). The specific implementation of the algorithm is described by Too et al. (2018). The mathematical equations of the models developed for the hunting strategies of wolves are given in Equations 5 and 6:

$$X(t+1) = X_p(t) - A.D \quad (5)$$

Where; X_p is the position of the prey, A is the coefficient vector, and D is defined as:

$$D = \left| C.X_p(t) - X(t) \right| \quad (6)$$

Where; C is the coefficient vector, X is the position of the grey wolf.

The position updates of the grey wolves take place as in Equation 7:

$$X(t+1) = \frac{X_1 + X_2 + X_3}{3} \quad (7)$$

2.3.3. Binary harris hawks optimization

Harris hawks optimization (HHO) algorithm is a population-based algorithm. It is a new swarm intelligence optimization algorithm inspired by the behavior and hunting patterns of Harris hawks, referred to as "surprise attacks". Harris hawks are one of the most intelligent hunting birds known. When a group of hawks get together and start the hunt, some of them make short tours one after the other and then descend into very high turnstiles. In this strategy, the hawks detect and attack their prey from different directions and approach simultaneously (Heidari 2019).

There are two different methods in the hunting process of HHO that decide which method will be used according to the randomly generated "q" value between [0-1]. In addition, a random number "r" is assigned between [0-1]. Different strategies are applied according to this "r" value and "E" escape energy. The "E" escape energy of the prey determines the attack on the prey. There are four different methods of producing solutions including soft besiege, hard besiege, developing attacks and soft besiege, and developing attacks and hard besiege (Çelik et al. 2019).

In Equation (8) and equation (10), the motion position equation and escape energy are defined for the new solution (Zhang et al. 2021):

$$x(t+1) = \begin{cases} x_r(t) - r_1.|x_r(t) - 2r_2 x(t), & q \geq 0.5 \\ (x_{target}(t) - x_{average}(t)) - r_3\,(r_4(UB - LB) + LB), & q < 0.5 \end{cases} \quad (8)$$

X (t + 1) and X (t) are the position vectors of the search agents. q, r1, r2, r3, and r4 are random values in each iteration and are randomly generated in the range 0-1. Xr(t) represents the position vector of a random individual. X_{target} (t) is the position vector of the prey. UB and LB show the lower and upper limits of the variables. $X_{average}$ (t) in Equation 3 is the average position vector of the available search agents, which can be calculated as (Zhang et al. 2021).

$$X_{average}(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t) \quad (9)$$

N is the population size of the hawks; $X_i(t)$ represents the position of an individual moving towards the prey (Zhang et al. 2021).

$$Escaping\_energy = 2E_0\left(1 - \frac{t}{T}\right) \quad (10)$$

In Equation 10 escape energy is defined for the new solution where T is the maximum number of iterations, E_0 is the initial energy value (Zhang et al. 2021). kNN error rate was used as the fitness function of this algorithm.

2.3.4. Binary particle swarm optimization algorithm

Particle Swarm Optimization (PSO) is a meta-heuristic algorithm that was inspired by the movements of swarms of animals, such as birds and fish (Kennedy & Eberhart 1995). It uses two important parameters, known as pbest and gbest, to update the velocity and position information of the candidate solutions in the swarm. The pbest value represents the local best solution, while the gbest value represents the global best solution.

The calculations of the algorithm are given in Equations 11-15 (Too et al. 2019):

$$v_i^d(t+1) = wv_i^d(t) + c_1 r_1\left(pbest_i^d(t) - x_i^d(t)\right) + c_2 r_2\left(gbest^d(t) - x_i^d(t)\right) \quad (11)$$

$$S\left(v_i^d(t+1)\right) = \frac{1}{1+\exp(-v_i^d(t+1))} \quad (12)$$

$$x_i^d(t+1) = \begin{cases} 1, & if\ rand < S\left(v_i^d(t+1)\right) \\ 0, & otherwise \end{cases} \quad (13)$$

Where; rand is a random number uniformly distributed between 0 and 1:

$$pbest_i^d(t+1) = \begin{cases} x_i(t+1), & if\ F\left(x_i(t+1)\right) < F(pbest_i(t)) \\ pbest_i(t), & otherwise \end{cases} \quad (14)$$

$$gbest(t+1) = \begin{cases} pbest_i(t+1), & if\ F\left(pbest_i(t+1)\right) < F(gbest(t)) \\ gbest(t)\ , otherwise \end{cases} \quad (15)$$

Where; x is the solution, pbest is personal best and gbest is global best solution. F(.) is fitness function. t is number of iterations.

BPSO is the binary version of the PSO algorithm. In this study, the following parameter values were used: c1=2; c2=2; Vmax=6; Wmax=0.9; Wmin=0.4.

2.4. Classification and evaluation of selected features

The classification process involves two phases: training (80% of the data) and testing (20% of the data). The dataset, consisting of features selected through the classification process, is divided into training and testing sets using cross-validation with a k value of 5. In the training phase, the parameters of the classification model are set and the resulting error is used to assess how well the model fits the training data. The testing phase demonstrates the model's ability to accurately predict labels for untested data. In this study, the kNN, SVM, and decision tree classification algorithms, which have previously been used to predict *Salmonella* in agricultural waters using the same dataset, were chosen to evaluate the performance of the meta-heuristic feature selection optimization (Polat et al. 2020; Buyrukoğlu 2021).

Figure 1 illustrates the four possible output states, representing the elements of a 2x2 confusion matrix or contingency table. The blue diagonal represents correct predictions, while the yellow diagonal indicates incorrect predictions. If a sample is positive and classified as positive, it is counted as a true positive (TP). If it is classified as negative, it is considered a false negative (FN). If a sample is negative and classified as negative, it is considered a true negative (TN). If it is classified as positive, it is considered a false positive (FP) (Tharwat 2018).

	True/Actual Class	
	Positive	Negative
Predicted Class	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Figure 1- An illustrative example of the 2X2 confusion matrix with two classes of Positive and Negative for classification. The output of the predicted class is defined as true or false

One of the most commonly used measures for evaluating classification performance is accuracy, which is calculated as the ratio of correctly classified samples to the total number of samples (Eq. 16). The precision, recall, and F-score metric values are given in Equations 17-19, respectively:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$\text{Micro Average Precision} = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} FP_k} \quad (17)$$

$$\text{Micro Average Recall} = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} FN_k} \quad (18)$$

$$\text{Micro Average F} - \text{score} = \frac{\sum_{k=1}^{K} TN_k}{\sum_{k=1}^{K} FP_k} \quad (19)$$

The complement of the accuracy metric is the error rate or misclassification rate, which reflects the number of misclassified samples from both positive and negative classes (Bradley 1997). It is calculated as follows (Eq. 20):

$$Error = 1 - Accuracy = (FP + FN)/(TP + TN + FP + FN) \quad (20)$$

This metric can be expressed as a percentage by multiplying the result by 100.

The micro-average score was used when equal weighting was required for each sample or estimate. The micro-average sums the contributions of all classes to calculate the average metric. In general, 'micro' is preferred where greater emphasis is placed on accuracy. For this reason, the micro-average was preferred in this study.

2.4.1. k-Nearest-neighbours classification

The k-Nearest-Neighbors (kNN) classifier is a simple and effective non-parametric classification method (Hand et al. 2001). In the kNN algorithm, the first step is to determine the distance between the data points. Common methods for measuring distance include the Euclidean, Manhattan, and Minkowski methods.

The Euclidean distance method, most commonly used in practice, is defined between samples Xi and Xj as shown in Equation 21:

$$(X_i, X_j) = \sqrt{(X_{i1} - X_{j1)}{}^2 + (X_{i2} - X_{j2)}{}^2 + \cdots + (X_{in} - X_{jn)}{}^2} \quad (21)$$

Another important factor in the kNN algorithm is the k parameter, which determines the number of neighboring values to consider when classifying a point. Selecting an appropriate k value is crucial for the success of the classification (Guo et al. 2003). To determine the best k value, the algorithm is run with different values of k and the performance is evaluated. A small value of k may result in too many classes, while a large value may lead to fewer classes than necessary and higher error rates (Imandoust & Bolandraftar 2013). In this study, the k value was set to 5 for each variable (k=5). The fitcknn toolbox in Matlab uses the Euclidean distance as the default distance measure.

2.4.2. Support vector machine classification

Support Vector Machines (SVMs) are a type of algorithm used for pattern recognition and classification tasks (Cortes & Vapnik 1995). They are based on statistical learning theory and are known for their ability to achieve good generalization performance. SVMs are particularly useful for dealing with large data sets because they transform the classification problem into a squared optimization problem, which allows for faster solution times compared to other techniques (Osowski et al. 2004). Additionally, SVMs have been shown to have superior classification performance, computational complexity, and usability compared to other methods due to their optimization-based procedure (Nitze et al. 2012). The aim of SVMs is to find the optimal hyperplane that separates different classes by maximizing the distance between the support vectors of different classes (Ayhan & Erdoğmuş 2014).

In this study, the kernel, degree, and C parameters were used in the SVM algorithm. The kernel parameter determines the type of hyperplane used, with options including linear, rbf, sigmoid, and poly for nonlinear hyperplanes. The degree parameter controls the flexibility of the decision boundaries, with higher degrees allowing for more complex nonlinear relationships between the features.

Equations (22) and (23) represent formulas for a line or hyper plane, respectively. The SVM should find weights so that the data points are separated according to a decision rule.

$$wx+b=0 \quad (22)$$

$$y=mx+b \quad (23)$$

The C parameter controls the trade-off between minimizing misclassifications and maximizing the margin between the classes. Higher values of C result in a tighter margin and fewer misclassifications, while lower values allow for more overlap between the classes and prioritize maintaining a maximum margin. The Matlab fitcsvm toolbox defaults to C values in the range of [0 1;1 0].

2.4.3. Decision tree classification

Decision trees are a type of machine learning algorithm used for building classifiers. They consist of decision nodes, which represent tests on a single attribute, and leaf nodes, which represent the resulting class. In binary decision trees, each decision node has two branches, one for each possible outcome of the attribute test. There are several decision tree algorithms, including ID3 (Iterative Dichotomiser 3), C4.5, CART (Classification and Regression Tree), CHAID (CHi-squared Automatic Interaction Detector), QUEST (Quick, Unbiased, Efficient, Statistical Tree), and MARS. In this study, the fitctree (binary decision trees) tool in Matlab was used.

The process of constructing a decision tree involves dividing the training data into smaller subsets and repeating this process until each subset belongs to a single class. The training data, represented by T, consists of k classes. If T consists of only one

class, it will be a leaf node. If T contains more than one class, it is divided into n subsets, where n is the number of outcomes for the attribute test ai. This process is repeated iteratively on each subset Tj (1<j<n) until each subset belongs to a single class (Buyrukoğlu et al. 2021). The default parameters of the Matlab fitctree toolbox used in this study were: MaxNumSplits = n-1, where n is the training sample size; MinLeafSize = 1; and MinParentSize = 10.

3. Results

3.1. Performance of meta-heuristic optimization algorithms for feature selection

Table 1 shows the feature distribution at each 100 steps for the binary version of DEO, GWO, HHO, and PSO optimization algorithms, respectively. The proposed feature selection approach was run 100 times for each meta-heuristic algorithm. The values in the Table 1 show how many times the relevant feature value is ranked among the first five features. However, when looking at the sum of some features, the total value appears below 100. This is because the algorithm discovers some parameters out of the first five features. Since there is an inherent randomness in heuristic algorithms, it is expected that such situations can occur.

All microbiological variables have been selected at least as the first, second, and third features. Total coliform was ranked as the first feature or non-feature among the selected first five features by all algorithms. HHO and PSO optimization chose total coliform 41 and 55 out of 100 times as the first feature for the prediction of *Salmonella* occurrences in agricultural water. Generic *E. coli* was selected as the first feature 81 and 91 times by GWO and DEO, respectively. However, HHO and PSO algorithms ranked generic *E. coli* 31 and 54 times as the first or second effective feature for prediction, respectively. *Enterococci* was chosen the highest 31 out of 100 times by all algorithms as the first or second feature. The GWO algorithm determined only microbiological indicators as the first feature followed by DEO (97 times), PSO (93 times), and HHO (69 times) for prediction of *Salmonella* in agricultural waters.

Air and water temperature were determined as the highest selected second and third features by all tested meta-heuristic algorithms with a selection range from 3 to 50 times. Conductivity, pH and oxidation-reduction potential of agricultural waters were ranked the highest as third, fourth and fifth features except for conductivity selected by the HHO algorithm. Turbidity and rain had the highest performance as fourth and fifth ranked features with the number of selections times below 19. The rest of the features were not consistently chosen as the successful feature for the prediction of *Salmonella* occurrence in agricultural waters. All meta-heuristic algorithms did not rank the rest of the features as the first or second feature, with the exception of 60 cm air temperature by the HHO algorithm.

Table 1- Numbers of selected first five features by tested meta-heuristic optimization algorithms for prediction of *Salmonella* occurrence in agricultural waters with classifiers

	Meta-heuristic Optimization Algorithms																			
	Binary Differential Evolution Optimization					*Binary Grey Wolf Optimization*					*Binary Harris Hawks Optimization*					*Binary Particle Swarm Optimization*				
Feature	*1st*	*2nd*	*3rd*	*4th*	*5th*	*1st*	*2nd*	*3rd*	*4th*	*5th*	*1st*	*2nd*	*3rd*	*4th*	*5th*	*1st*	*2nd*	*3rd*	*4th*	*5th*
Total Coliform	4	0	0	0	0	15	0	0	0	0	41	0	0	0	0	55	0	0	0	0
Generic *E. coli*	91	2	0	0	0	81	5	0	0	0	9	22	0	0	0	28	26	0	0	0
Enterococci	2	2	0	0	0	4	28	4	0	0	19	14	9	0	0	10	31	15	0	0
Air Temperature	0	50	3	0	0	0	42	21	3	0	14	19	10	6	0	3	19	19	5	0
Water Temperature	2	24	28	1	0	0	16	27	14	0	9	13	16	8	2	1	10	21	15	1
Conductivity	0	9	23	8	1	0	6	19	21	10	8	21	13	8	6	1	5	13	13	10
pH	1	8	15	30	3	0	1	18	19	13	0	2	15	15	5	1	2	12	20	15
Oxidation-Reduction Potential	0	3	11	17	24	0	2	2	22	24	0	1	11	11	11	1	2	8	14	17
Turbidity	0	2	8	15	19	0	0	8	9	18	0	2	2	9	8	0	4	2	9	11
Rain	0	0	4	10	11	0	0	1	5	15	0	0	2	3	12	0	1	5	9	10
Total Solar Radiation	0	0	7	9	21	0	0	0	7	9	0	0	1	3	6	0	0	0	4	7
Average Solar Radiation	0	0	0	3	8	0	0	0	0	6	0	0	14	9	4	0	0	1	4	9
60 cm Air Temperature	0	0	0	4	3	0	0	0	0	2	0	6	5	13	13	0	0	3	3	7
Relative humidity	0	0	1	1	2	0	0	0	0	0	0	0	1	5	9	0	0	1	2	5
Ten-meter Wind Direction	0	0	0	1	4	0	0	0	0	2	0	0	1	0	1	0	0	0	1	3
Ten-meter Wind Speed	0	0	0	1	2	0	0	0	0	1	0	0	0	6	7	0	0	0	1	3
60 cm Soil Temperature	0	0	0	0	1	0	0	0	0	0	0	0	0	3	3	0	0	0	0	0
Total Iteration	100	100	100	100	99	100	100	100	100	100	100	100	100	99	87	100	100	100	100	98

3.2. Frequency of meta-heuristic optimization algorithms for feature selection

Table 2 shows the feature selection frequency of meta-heuristic algorithms based on the first five ranked features. Generic *E. coli* was ranked as the most successful feature in the prediction of *Salmonella* presence in agricultural waters ranging from 31 to 93 times selected in the first five ranked features. Similar selection frequency was observed for total coliform and enterococci. DEO was the only algorithm to rank almost all physicochemical attributes and weather station measurement above microbiological indicators. Air and water temperature were ranked between 46 and 66 times in the first five features by all meta-heuristic algorithms. Similar to results for microbiological indicators, the distribution of feature selection was parallel between GWO and DEO and between PSO and HHO algorithms for air and water temperatures. 60 cm air temperature measurements from weather station was not selected as frequent feature as actual air and water temperatures measured in the field for prediction of *Salmonella* occurrence in agricultural waters. With the exception of HHO, conductivity, the pH and oxidation-reduction potential of agricultural were ranked from 41 to 57 times in the first five features based on all meta-heuristic algorithms. Turbidity was ranked in the first five features 21 to 44 times by all algorithms. The rest of the features including rain and solar radiation were ranked between none and 37 times in the first five algorithms. When the feature selections of all meta-heuristic algorithms were combined, generic *E. coli* (264 times) was selected almost twice as often as total coliform (115 times) and enterococci (138 times). The order of feature dominancy was generic *E. coli* (264 times), air temperature (214 times), water temperatures (208 times), conductivity and pH (195 times).

Table 2- Total number of features selected as first and second feature by four tested meta-heuristic optimization algorithms for prediction of *Salmonella* occurrence in agricultural waters with classifiers

	*Frequency**				
Feature	*DEO*	*GWO*	*HHO*	*PSO*	*Total*
Total Coliform	4	15	41	55	115
Generic *E. coli*	93	86	31	54	264
Enterococci	4	36	42	56	138
Air Temperature	53	66	49	46	214
Water Temperature	55	57	48	48	208
Conductivity	41	56	56	42	195
pH	57	51	37	50	195
Oxidation-Reduction Potential	55	50	34	42	181
Turbidity	44	35	21	26	126
Rain	25	21	17	25	88
Total Solar Radiation	37	16	10	11	74
Average Solar Radiation	11	6	27	14	58
60 cm Air Temperature	7	2	37	13	59
Relative humidity	4	0	15	8	27
Ten-meter Wind Direction	5	2	2	4	13
Ten-meter Wind Speed	3	1	13	4	21
60 cm Soil Temperature	1	0	6	0	7

*: Binary meta-heuristic optimization methods include differential evolution optimization (DEO), grey wolf optimization (GWO), Harris hawks optimization (HHO) and particle swarm optimization (PSO).

3.3. Performance of classifier based on selected meta-heuristic algorithm

Table 3 depicts the prediction accuracy results of kNN, SVM, and decision tree based on feature selection by tested meta-heuristic algorithms. Since each meta-heuristic method was run for 100 steps, the classification process was repeated each time to obtain the average and the highest prediction success as a percentage. All classification algorithms predicted the *Salmonella* occurrence based on selected features with accuracy values ranging from 93.70% to 95.18% on average. The highest accuracy rates based on feature selection by all meta-heuristic algorithms were 95.18% for kNN, 95.16% for SVM, and 95.55% for decision tree. The average accuracies predicted by SVM and KNN were 95.18% for all meta-heuristic algorithms. The average accuracy success rates of decision tree ranged from 93.70% to 95.55%.

Similar to accuracies, other evaluation parameters including precision, recall and f-score were calculated over 93.00% regardless of the feature selection and classification algorithm. The highest precision value was obtained from the kNN and SVM algorithms with 95.18%. The average precision value in the decision tree classification in all algorithms was lower than the other classifiers. The same results were observed in metrics such as recall, f-score that were between 93.57 and 95.18%. The kNN, SVM and DT accuracy results are shown in Figure 2 for each meta-heuristic algorithm (DEO, GWO, HHO and PSO).

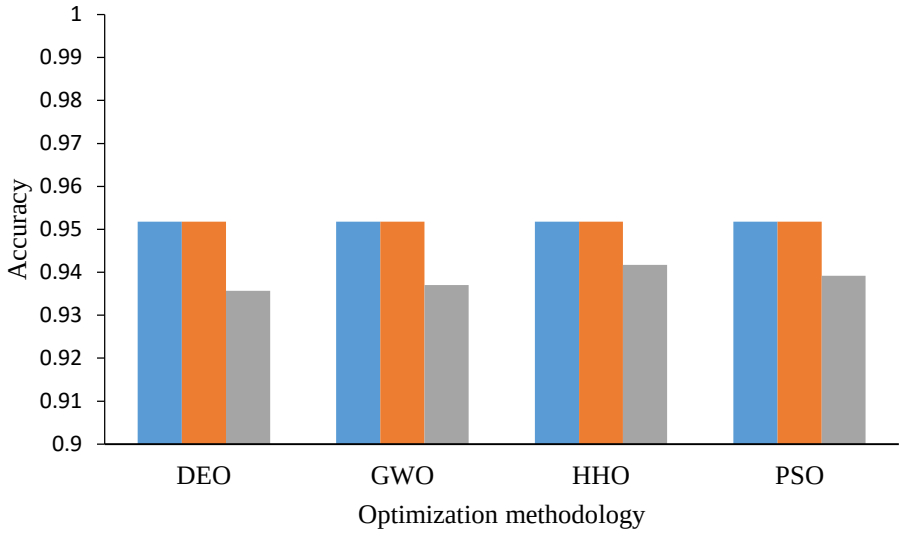

Figure 2- Accuracy results of kNN (■), SVM (■) and DT (■) classifiers after application of differential evolution optimization (DEO), grey wolf optimization (GWO), Harris hawks optimization (HHO) and particle swarm optimization (PSO) heuristics algorithms for feature selection

Table 3- Accuracy results of kNN, SVM, and decision tree (DT) classifications based on feature selection of meta-heuristic optimization algorithms

Algorithm	*Classification*	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F-Score (%)*
DEO	kNN-Maximum	0.9518	0.9518	0.9518	0.9518
	kNN-Average	0.9518	0.9518	0.9518	0.9518
	SVM-Maximum	0.9518	0.9518	0.9518	0.9518
	SVM-Average	0.9518	0.9518	0.9518	0.9518
	DT-Maximum	0.9481	0.9481	0.9481	0.9481
	DT-Average	0.9357	0.9357	0.9357	0.9357
GWO	kNN-Maximum	0.9518	0.9518	0.9518	0.9518
	kNN-Average	0.9518	0.9518	0.9518	0.9518
	SVM-Maximum	0.9518	0.9518	0.9518	0.9518
	SVM-Average	0.9518	0.9518	0.9518	0.9518
	DT-Maximum	0.9555	0.9555	0.9555	0.9555
	DT-Average	0.9370	0.9370	0.9370	0.9370
HHO	kNN-Maximum	0.9518	0.9518	0.9518	0.9518
	kNN-Average	0.9518	0.9518	0.9518	0.9518
	SVM-Maximum	0.9518	0.9518	0.9518	0.9518
	SVM-Average	0.9518	0.9518	0.9518	0.9518
	DT-Maximum	0.9518	0.9518	0.9518	0.9518
	DT-Average	0.9417	0.9417	0.9417	0.9417
PSO	kNN-Maximum	0.9518	0.9518	0.9518	0.9518
	kNN-Average	0.9518	0.9518	0.9518	0.9518
	SVM-Maximum	0.9518	0.9518	0.9518	0.9518
	SVM-Average	0.9518	0.9518	0.9518	0.9518
	DT-Maximum	0.9499	0.9499	0.9499	0.9499
	DT-Average	0.9392	0.9392	0.9392	0.9392

*: Binary meta-heuristic optimization methods include differential evolution optimization (DEO), grey wolf optimization (GWO), Harris hawks optimization (HHO) and particle swarm optimization (PSO)

4. Discussion

The microbiological water quality via indicator microorganisms is monitored to the reduce pathogen contamination risk of produce. The detection of pathogens is also possible for agricultural water, but not preferred due to the high cost, length of time, advanced laboratory equipment and qualified personnel required to perform any analysis. To provide faster and easier risk assessment for growers, several statistical and computer-based approaches have been proposed for the prediction of pathogen occurrence in agricultural waters over the past (Benjamin et al. 2013; McEgan et al. 2013; Bradshaw et al. 2016; Havelaar et al. 2017; Truchado et al. 2018; Polat et al. 2020; Weller et al. 2020; Buyrukoğlu 2021;). The success of each model or algorithm varies depending on intrinsic and extrinsic parameters used for prediction. Several features including the population of microbiological indicators, physicochemical attributes, or environmental variables can be used for the prediction of a *Salmonella* population or occurrence in agricultural water sources (Buyrukoğlu 2021). However, unstable environmental conditions may dramatically affect variable changes used as features and, relatively, the performance of prediction tools. Preprocessing of parameter values can be a requirement for the success of prediction. In this study, since the intervals of the values in the data set contained various ranges and units, feature values were subjected to the normalization process before optimization.

The dataset used for the prediction of *Salmonella* presence in agricultural waters included 17 possible features with various ranges and units in this study. The order of feature dominancy was generic *E. coli*, air temperature, water temperatures, conductivity and pH by tested meta-heuristic algorithms; however, 60 cm air temperature was determined as a weak predictor. This is because air and water temperatures were measured on-site while the 60 cm air temperature data was taken from the weather station (Topalcengiz et al. 2017). The meta-heuristic algorithms used for optimization ranked indicator microorganisms in the first and second place of the most selected features among the first five features. Particularly, the generic *E. coli* population was chosen as the first feature except for the HHO algorithm. Previously, the same dataset was evaluated for the best *Salmonella* prediction in agricultural waters with statistical and computer-based tools. Havelaar et al. (2017) developed a prediction model for the probability of the presence of *Salmonella* by using the *E. coli* population and turbidity based on the results of logistic regression analysis for feature selection. In another study with the same dataset, heterogenous ensemble feature selection including information gain, RelifF, analysis of variance, and Chi-square yielded the most successful *Salmonella* prediction with features including UV, turbidity, and the population of coliform and generic *E. coli* (Buyrukoğlu 2021). In the same study, microbiological indicators are noted as more effective features than physicochemical attributes and weather station measurements as meta-heuristic algorithms used in here.

Feature selection aims to find a subset of features for a learning operation that can describe data as well or better than the original dataset (Phyu & Oo 2016). In feature selection, there are three groups: filtering methods based on statistical information, spiral search methods, and embedded methods using the best divisor criterion. In filtering methods, feature selection is made before the selection algorithm works, while in spiral methods, the algorithm is used for the selection of the best features. In embedded methods, the data mining algorithm and feature selection algorithm work simultaneously (Budak 2018). In this study, four meta-heuristic methods were used for filtering before classification with kNN, SVM, and decision tree algorithms for the prediction of *Salmonella* occurrence in light of previous studies analyzing the same data set (Buyrukoğlu 2021; Buyrukoğlu et al. 2021; Polat et al. 2020). This study can be considered as the performance of an ablation study in the analysis of agricultural water quality.

Generic *E. coli* was ranked as the first feature (and dominant feature in total) over 80 out of 100 times by GWO and DEO for the prediction of *Salmonella* presence or absence in agricultural waters as previous studies conducted with the same data set (Buyrukoğlu 2021; Polat et al. 2020). However, HHO and PSO algorithms listed *E. coli* the highest 28 times as the first or second among the first five features. These differences show that swarm or population size-based optimization algorithms are not as successful as continuous optimization algorithms as GWO and DEO.

Previously, Polat et al. (2020) reported the highest accuracy around 76% with ANN, kNN and SVM classifiers by using the same dataset as individual or combined features. In their study, no feature selection was performed among microbiological indicators or physiochemical water attributes. In another study with the same dataset, Buyrukoğlu (2021) proposed a new hybrid data mining model for prediction of *Salmonella* occurrence. Ensemble models of ANN, SVM, random forest and Naïve Bayes using a heterogenous feature selection approach (information gain, RelifF, analysis of variance, and Chi-square) had a prediction accuracy ranging from 82.8 to 94.9% (Buyrukoğlu 2021). In this study, the accuracy success of kNN, SVM, and decision tree algorithms was determined between 93.70 and 95.55% on average after 100 iterations based on tested metaheuristic optimization algorithms. High accuracy calculations show that the feature selection obtained through the use of heuristic methods yields more successful results.

The method proposed in this study gave higher prediction accuracy results than previous studies using the same data set. Recently, Buyrukoğlu et al. (2022) managed to increase the prediction success of the deep feed-forward neural network (DFNN) for *Salmonella* occurrence up to an accuracy of 98.41% with determined correlation value based on the selected features in another study with the same dataset. In Buyrukoğlu's (2022) study, feature selection determined by gain ratio yielded the highest relationships between generic *E. coli* and rain, solar radiation, and turbidity. Then, predicted generic *E. coli* population by decision tree, SVM, and RF were combined with selected environmental and physicochemical features with and without

correlation value for the DFNN analysis (Buyrukoğlu et al. 2022). The generic *E. coli*, air temperature, water temperature, conductivity and pH selected by meta-heuristic methods appear to be more dominant in the prediction of *Salmonella* presence in agricultural waters. In addition, as a result of feature selection made with heuristic methods in here unlike previous studies using the same data set (Polat et al. 2020; Buyrukoğlu 2021; Buyrukoğlu et al. 2021; Buyrukoğlu et al. 2022), conductivity and pH stand out as advantageous and distinguishing features that can be measured with portable equipment in the field. The measured conductivity and pH values can be used with a computer-based tool or developed application to obtain immediate results.

A small number of positive *Salmonella* samples in our dataset can be considered as a limitation of this study due to imbalanced classifications. To overcome this limitation, the micro-average method (Grandini et al. 2020) was used to calculate the metrics in the classification process. The unique aspect of our study is that it is the first study in which feature selection was made using meta-heuristic algorithms on a dataset for the prediction of *Salmonella* in agricultural waters. At all combinations of feature selection and classification, prediction success was calculated higher than accuracies calculated with the same or similar datasets.

5. Conclusions

In this study, the data collected from six different agricultural ponds were analyzed. A classification study was conducted to predict the presence/absence of the *Salmonella* pathogen. In the first part of the proposed method, a feature selection was performed with four different meta-heuristic algorithms. Following this, a classification was made using the kNN, SVM and decision tree classification methods. Similar to previous studies using the same or similar data sets, generic *E. coli* was selected as the most prominent feature for the prediction of *Salmonella* occurrence in agricultural waters. This confirms the validity of the recommended microbiological indicator compared to water attributes and weather station measurements. The accuracy success of the classifiers was improved up to 95% after feature selection using the metaheuristic optimization algorithms. There has yet to be a study using heuristic optimization methods for feature selection on the same data set and performed classification with these features. In this respect, this study shows that the use of heuristic methods may improve results in future studies in this area, especially in cases where the data size and the number of parameters is high. In this study the main strength of the proposed model is the use of a hybrid approach that combines feature selection and machine learning.

Acknowledgements

The authors also thank Selim Buyrukoğlu for his support and advice.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research was supported by Mus Alparslan University.

Author contributions

Data curation was obtained by Zeynal Topalcengiz. Conceptualization, formal analysis, and methodology were performed by Murat Demir and Murat Canayaz. Resources, software, supervision, writing – review & editing were organized by Zeynal Topalcengiz, Murat Demir and Murat Canayaz.

References

Abimbola O P, Mittelstet A R, Messer T L, Berry E D, Bartelt-Hunt S L & Hansen S P (2020). Predicting Escherichia coli loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern. The Science of the Total Environment 722: 137894. https://doi.org/10.1016/j.scitotenv.2020.137894

Akinola O O, Ezugwu A E, Agushaka J O, Zitar R A & Abualigah L (2022). Multiclass feature selection with metaheuristic optimization algorithms: a review. Neural Computing and Applications 34: 19751-19790. https://doi.org/10.1007/s00521-022-07705-4

Agrawal P, Abutarboush H F, Ganesh T & Mohamed A W (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019). IEEE Access 9: 26766-26791. https://doi.org/10.1109/ACCESS.2021.3056407

Ashbolt N, Grabow W O K & Snozzi M (2001). Indicators of microbial water quality. In: L Fewtrell & J Bartram (Eds.), Water Quality: Guidelines, Standards and Health, World Health Organization (WHO) IWA Publishing pp. 289-316

Ayhan S & Erdoğmuş Ş (2014). Kernel function selection for the solution of classification problems via support vector machines. Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi (In Turkish). Eskişehir Osmangazi University Journal of Economics and Administrative Sciences 9:175-201

Benjamin L, Atwill E R, Jay-Russell M, Cooley M, Carychao D, Gorski L & Mandrell R E (2013). Occurrence of generic Escherichia coli, E. coli O157 and Salmonella spp. in water and sediment from leafy green produce farms and streams on the Central California coast. International Journal of Food Microbiology 165(1): 65-76. https://doi.org/10.1016/j.ijfoodmicro.2013.04.003

Blum C & Roli A (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys 35: 268-308. https://doi.org/10.1145/937503.937505

Bradley A P (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30: 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Bradshaw J K, Snyder B J, Oladeinde A, Spidle D, Berrang M E, Meinersmann R J, Oakley B, Sidle R C, Sullivan K & Molina M (2016). Characterizing relationships among fecal indicator bacteria, microbial source tracking markers, and associated waterborne pathogen occurrence in stream water and sediments in a mixed land use watershed. Water Research 101: 498-509. https://doi.org/10.1016/j.watres.2016.05.014

Budak H (2018). Feature selection methods and a new approach. Özellik seçim yöntemleri ve yeni bir yaklaşım (In Turkish). Süleyman Demirel University Journal of Natural and Applied Sciences 22: 21-31. https://doi.org/10.19113/sdufbed.01653

Buyrukoğlu S (2021). New hybrid data mining model for prediction of Salmonella presence in agricultural waters based on ensemble feature selection and machine learning algorithms. Journal of Food Safety 41: 12903. https://doi.org/10.1111/jfs.12903

Buyrukoğlu G, Buyrukoğlu S & Topalcengiz Z (2021). Comparing regression models with count data to artificial neural network and ensemble models for prediction of generic Escherichia coli population in agricultural ponds based on weather station measurements. Microbial Risk Analysis 19: 100171. https://doi.org/10.1016/j.mran.2021.100171

Buyrukoğlu S, Yılmaz Y & Topalcengiz Z (2022). Correlation value determined to increase Salmonella prediction success of deep neural network for agricultural waters. Environmental Monitoring and Assessment 194: 373. https://doi.org/10.1007/s10661-022-10050-7

Canayaz M (2021). MH-COVIDNet: Diagnosis of COVID-19 using deep neural networks and meta-heuristic-based feature selection on X-ray images. BIomedical Signal Processing and Control 64: 102257. https://doi.org/10.1016/j.bspc.2020.102257

Centers for Disease Control and Prevention (CDC) (2007). Multistate outbreaks of Salmonella infections associated with raw tomatoes eaten in restaurants--United States, 2005-2006. MMWR. Morbidity and Mortality Weekly Report 56(35): 909–911.

Cortes C & Vapnik V (1995). Support-vector networks. Machine Learning 20: 273-297. https://doi.org/10.1007/BF00994018

Çelik Y, Yıldız İ & Karadeniz A T (2019). A brief review of metaheuristic algorithms improved in the last three years. European Journal of Science and Technology pp. 463-477. https://doi.org/10.31590/ejosat.638431

Das S & Suganthan P N (2011). Differential Evolution: A Survey of the State-of-the-Art. IEEE Transactions on Evolutionary Computation 15: 4-31. https://doi.org/10.1109/TEVC.2010.2059031

Dokeroglu T, Deniz A & Kiziloz H E (2022). A comprehensive survey on recent metaheuristics for feature selection. Neurocomputing 494: 269-296. https://doi.org/10.1016/j.neucom.2022.04.083

Emary E, Zawbaa H M & Hassanien A E (2016). Binary grey wolf optimization approaches for feature selection. Neurocomputing 172: 371-381. https://doi.org/10.1016/j.neucom.2015.06.083

Food and Drug Administration (FDA) (2015). Federal Register Notice: Standards for the Growing, Harvesting, Packing, and Holding of Produce for Human Consumption; Final Rule. Available at: https://www.gpo.gov/fdsys/pkg/FR-2015-11-27/pdf/2015-28159.pdf. Accessed 12 July 2022

Grandini M, Bagli E & Visani G (2020). Metrics for Multi-Class Classification: An Overview. ArXiv, https://doi.org/10.48550/arXiv.2008.05756

Greene S K, Daly E R, Talbot E A, Demma L J, Holzbauer S, Patel N J, Hill T A, Walderhaug M O, Hoekstra R M, Lynch M F & Painter J A (2008). Recurrent multistate outbreak of Salmonella Newport associated with tomatoes from contaminated fields, 2005. Epidemiology and Infection 136(2): 157–165. https://doi.org/10.1017/S095026880700859X

Guo G, Wang H, Bell D, Bi Y & Greer K (2003). KNN model-based approach in classification. In: R Meersman et al (Eds.), On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, Springer, pp. 986-996. https://doi.org/10.1007/978-3-540-39964-3_62

Hand D, Mannila H & Smyth P (2001). Principles of data mining. A Bradford Book the MIT Press.

Havelaar A H, Vazquez K M, Topalcengiz Z, Muñoz-Carpena R & Danyluk M D (2017). Evaluating the U.S. Food Safety Modernization Act Produce Safety Rule standard for microbial quality of agricultural water for growing produce. Journal of Food Protection 80: 1832-1841. https://doi.org/10.4315/0362-028X.JFP-17-122

Heidari A A, Mirjalili S, Faris H, Aljarah I, Mafarja M & Chen H (2019). Harris hawks optimization: Algorithm and applications. Future Generation Computer Systems 97: 849-872. https://doi.org/10.1016/j.future.2019.02.028

Imandoust S B & Bolandraftar M (2013). Application of K-nearest neighbor (KNN) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications 3: 605-610.

Kennedy J & Eberhart R (1995). Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 4: 1942-1948. https://doi.org/10.1109/ICNN.1995.488968

Liang Y, Liao B & Zhu W. (2017). An improved binary differential evolution algorithm to infer tumor phylogenetic trees. BioMed Research International 2017: 5482750. https://doi.org/10.1155/2017/5482750

McEgan R, Mootian G, Goodridge L D, Schaffner D W & Danyluk M D (2013). Predicting Salmonella populations from biological, chemical, and physical indicators in Florida surface waters. Applied and Environmental Microbiology 79(13): 4094-4105. https://doi.org/10.1128/AEM.00777-13

Mirjalili S, Mirjalili S M & Lewis A. (2014). Grey wolf optimizer. Advances in Engineering Software 69: 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007

Nitze I, Schulthess U & Asche H (2012). Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. Proceedings of the 4th GEOBIA 35-40.

Osowski S, Siwek K & Markiewicz T (2004). MLP and SVM networks - a comparative study. Proceedings of the 6th Nordic Signal Processing Symposium pp. 37-40

Phyu T Z & Oo N N (2016). Performance comparison of feature selection methods. MATEC Web of Conferences 42: 06002. https://doi.org/10.1051/matecconf/20164206002

Polat H, Topalcengiz Z & Danyluk M D (2020). Prediction of Salmonella presence and absence in agricultural surface waters by artificial intelligence approaches. Journal of Food Safety 40: e12733. https://doi.org/10.1111/jfs.12733

Price K V, Storn R M & Lampinen J A (2005). Differential evolution: A practical approach to global optimization, Springer https://doi.org/10.1007/3-540-

Steele M, Mahdi A & Odumeru J (2005). Microbial assessment of irrigation water used for production of fruit and vegetables in Ontario, Canada. Journal of Food Protection 68(7): 1388–1392. https://doi.org/10.4315/0362-028X-68.7.1388

Storn R & Price K (1997). Differential evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces. Journal of Global Optimization 11: 341-359. https://doi.org/10.1023/A:1008202821328

Tharwat A (2018). Classification assessment methods. Applied Computing and Informatics 17: 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Too J, Abdullah A R, Mohd Saad N M, Ali N M & Tee W (2018). A new competitive binary grey wolf optimizer to solve the feature selection problem in EMG signals classification. Computers 7: 58. https://doi.org/10.3390/computers7040058

Too J, Abdullah A R, Mohd Saad N M & Tee W (2019). EMG feature selection and classification using a Pbest-guide binary particle swarm optimization, Computation 7(1): 12. https://doi.org/10.3390/computation7010012

Topalcengiz Z & Danyluk M D (2019). Fate of generic and Shiga toxin-producing Escherichia coli (STEC) in Central Florida surface waters and evaluation of EPA Worst Case water as standard medium. Food Research International 120: 322-329. https://doi.org/10.1016/j.foodres.2019.02.045

Topalcengiz Z, McEgan R & Danyluk M D (2019). Fate of Salmonella in Central Florida surface waters and evaluation of EPA Worst Case Water as a standard medium. Journal of Food Protection 82(6): 916–925. https://doi.org/10.4315/0362-028X.JFP-18-331

Topalcengiz Z, Strawn L K & Danyluk M D (2017). Microbial quality of agricultural water in Central Florida. PLoS ONE 12(4): e0174889. https://doi.org/10.1371/journal.pone.0174889.

Truchado P, Hernandez N, Gil M I, Ivanek R & Allende A (2018). Correlation between E. coli levels and the presence of foodborne pathogens in surface irrigation water: Establishment of a sampling program. Water Research 128: 226–233. https://doi.org/10.1016/j.watres.2017.10.041

Weller D L, Love T, Belias A & Wiedmann M (2020). Predictive Models may complement or provide an alternative to existing strategies for assessing the enteric pathogen contamination status of northeastern streams used to provide water for produce production. Frontiers in Sustainable Food Systems 4: 561517. https://doi.org/10.3389/fsufs.2020.561517

Yang X S (2011). Review of metaheuristics and generalized evolutionary walk algorithm. International Journal of Bio-Inspired Computation 3: 77-84. https://doi.org/10.1504/IJBIC.2011.039907

Zhang Y, Liu R, Wang X, Chen H & Li C (2021). Boosted binary Harris hawks optimizer and feature selection. Engineering with Computers 37: 3741-3770. https://doi.org/10.1007/s00366-020-01028-5