International Journal of Mathematics and
Computer Science, **19**(2024), no. 4, 1005–1018

Advanced Machine Learning Swarm Intelligence Algorithms in Atmospheric Pollutants Prediction

Somia Asklany, Salwa Othmen, Wahida Mansouri

Department of Computer Science and Information Technology
Faculty of Sciences and Arts, Turaif
Northern Border University
Arar 91431, Kingdom of Saudi Arabia

email: somiaasklany@hotmail.com

(Received February 11, 2024, Accepted March 19, 2024,
Published June 1, 2024)

Abstract

Advanced Machine learning based on the combination of an Artificial Bee Colony (ABC) intelligent algorithm with a "least square support vector machine" (LSSVM) is used to predict hourly values for three air pollutants (Carbon monoxide (CO), Sulphur Dioxide (S2O), and Nitrogen Dioxide(N2O)). ABC is explored using a stochastic population-based evolutionary algorithm to tackle complex problems. This algorithm is regarded as one of the most recent advancements in swarm intelligence techniques and was combined with the optimization of the least square support vector machine (LSSVM) for constructing predictive models for atmospheric pollutants. The proposed model is built based on long-term recorded hourly dataset for New Cairo station Cairo city coordinates $30°14'0''\text{N}, 31°12'22\text{E}$ for three years 2020-2022 and this huge data set was divided into train and test sets to develop the model through select through optimizing LSSVM by The ABC algorithm was employed to select the optimal set of free parameters for LSSVM, aiming to mitigate issues such as overfitting, local minima, and enhance the precision of prediction models.

Keywords and phrases: Machine Learning, Intelligence Algorithms
AMS (MOS) Subject Classifications: 68T01, 68T05, 68T27.
The corresponding author is Somia Asklany.
ISSN 1814-0432, 2024, http://ijmcs.future-in-tech.net

Multi-Layer Perceptron (MLP) is a comparison technique against the proposed model, both applied on check unseen dataset. The proposed model demonstrated high performance and efficiency compared with MLP by achieving better root mean square (RMSE) error and mean absolute error (MAE) than MLP.

1 Introduction

Pollution is one of the most significant problems of urban areas. With an increase in population and economic progress and a change in lifestyle leading to new industries, environmental health problems appear to capture society's interest. These Problems affect the ecosystem and can lead to climate change. Obviously, air pollution has a direct effect on people's health. Lung cancer, Asthma, ventricular hypertrophy, Parkinson's and Alzheimer's diseases, autism, psychological difficulties, retinopathy, prenatal development, and low birth weight are only a few of the conditions that are thought to be pointedly impacted by air pollution [1]. The most important pollutants are particulate matter (PM), sulfur dioxide (SO2), nitrogen dioxide (NO2), and carbon monoxide (CO). In Megacities like Cairo, where over 10 million people live, huge population expansion and unsustainable urban development have increased transportation, industrial activity, and energy demand, resulting in increased air pollution that exposes communities to serious hazards. The concentration of air pollutants, in particular gaseous pollutants, is increased by the burning of fossil fuels. Many applications are concerned with energy production, particularly in the industrial and transportation sectors [2]. Air quality is a major determinant of environmental health and public welfare, and forecasting air pollutant concentrations plays a key role in reducing and mitigating their negative impacts. Traditional techniques, such as statistical methods used to predict air pollutants. However, these methods often face challenges related to the complexity of atmospheric processes, high data dimensionality, and the need for real-time predictions. In this context, prediction tools for pollutant concentrations are established: Autoregressive Integrated Moving Average ARIMA is a commonly used method in time series prediction, and The Generalized Autoregressive Conditional Heteroskedasticity (GARCH), which is a conventional method primarily employed for linear time series forecasting, is also utilized [3]. Machine learning classification techniques prove great ability in prediction processes [4]. The Artificial Neural Networks (ANN) method, as one of the most important branches of Artificial Intelligence (AI), is a common machine learning technique used in

massive nonlinear prediction problems. The utilization of Artificial Neural Networks (ANNs) for prediction aims to address the constraints of conventional techniques. However, ANNs frequently confront the challenge of overfitting, primarily stemming from the extensive parameter set that needs to be adjusted and a lack of prior developer knowledge regarding the significance of input vectors during problem analysis [5]. Support vector machines (SVMs) have been devised as a substitute, circumventing the constraints of ANN prediction models. Their practical achievements can be ascribed to robust theoretical underpinnings rooted in the Vapnik- Chervonenkis (VC) theory [6]. SVM calculates globally optimal solutions, unlike ANN, which occasionally converges to local minima, resulting in dissimilar outcomes [7]. The Least Squares Support Vector Machine (LSSVM) approach offers a revamped rendition of the traditional SVM algorithm. LSSVM utilizes a regularized least squares objective function with equality constraints to formulate a linear system that adheres to the Karush-Kuhn-Tucker (KKT) conditions, achieving an optimal solution. Despite LSSVM simplifying the SVM procedure, the primary determinants influencing the performance of the regression system are the regularization and kernel parameters. Therefore, it is crucial to establish an effective methodology for selecting free parameters in LSSVM. This ensures that the resulting regression remains robust in the face of noisy data, eliminating the necessity for prior user knowledge regarding the impact of these parameters on the specific problem at hand [8]. The combination of machine learning and swarm intelligence algorithms has recently been recognized as a powerful and innovative approach to improving the accuracy and efficiency of air pollutant forecasting.

2 Methodology

This study introduces a hybrid model that combines LSSVM and ABC for predicting pollutants. The LSSVM's performance relies on the careful selection of hyperparameters, including C (cost penalty), ϵ (insensitive loss function), and γ (kernel parameter). ABC is employed to identify the optimal combination of these parameters for LSSVM. Detailed explanations of these concepts will be provided in the subsequent two sections.

2.1 Least Square Support Vector Machine

LSSVM classifiers proposed by Suykens and Vandewalle [9]. Least squares support vector machines (LSSVM) are seen as an alternative version of SVM.

They represent a variation of SVM that seeks to identify patterns by establishing a connection between input and output data, making them applicable for both classification and regression tasks. In contrast to traditional SVMs, which involve solving complex convex quadratic programming (QP) problems, LSSVM addresses the problem by solving a set of linear equations. If X is $n \times p$ matrix for input data and y is $n \times 1$ output vector. Consider the $\{x_i, y_i\}_{i=1}^{n}$ training data set of n data points, where $xi \in R^p$ and $y_i \in R$, the LSSVM target is to construct the function $f(x) = y$, which indicates the dependence of the output vector y_i on the input set x_i. This mapping function expressed as

$$f(x) = W^T \phi(x) + b, \tag{2.1}$$

where W an adjustable weight vector and $\phi(x)$ nonlinear mapping:

$R^p \to R^n$ are $n \times 1$ column vectors, and $b \in R$ is the scalar threshold. The key distinction in this criterion is that LSSVM replaces inequalities with equality constraints and employs a least square cost function. Additionally, the LSSVM approach tackles a linear problem, whereas the standard SVM deals with a quadratic one. Minimization optimization problem and the definition of equality constraints for LSSVM can be outlined as follows:

$$\begin{aligned} &\min_{w,e,b} j(w, e, b) = \frac{1}{2} w^T + C \frac{1}{2} e^T \\ &y_i = w^T \phi\left(x_i\right) + b + e_i \end{aligned} \tag{2.2}$$

Here, "e" represents a vector of errors with dimensions n × 1, "1" is an n × 1 vector with all elements set to 1 , and " C " ∈ $R^+$is the tradeoff parameter governing the balance between solution complexity and training errors. From equation (2.2), we derive a Lagrangian by taking differentials with respect to w, b, e, and a (where "a" signifies Lagrangian multipliers). We then have the following matrix equation:

$$\begin{bmatrix} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -1^T \\ 0 & 0 & CI & -I \\ Z & 1 & I & 0 \end{bmatrix} \begin{bmatrix} W \\ b \\ e \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ y \end{bmatrix} \tag{2.3}$$

where I represents the identity matrix and

$$Z = [\phi(x1), \phi(x2), \ldots, \phi\left(x_n\right)]^T. \tag{2.4}$$

From rows one and three in (2.3) $w = Z^T a$ and $Ce = a$

Then, by defining the kernel matrix $K = ZZ^T$, and the parameter $\lambda = C^{-1}$, the conditions to get an optimal solution is given by:

$$\begin{bmatrix} 0 & 1^T \\ 1 & K + \lambda I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{2.5}$$

where the Kernel function K types are defined as follows:
- Linear kernel given by:

$$K(x, x_i) = x_i^T x \tag{2.6}$$

- Polynomial kernel of degree specified by:

$$K(x, x_i) = \left(1 + x_i^T x/c\right)^d \tag{2.7}$$

- Radial basis function RBF kernel specified by:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\sigma^2}\right) \tag{2.8}$$

-Multi Layer Perceptron MLP kernel specified by:

$$K(x, x_i) = \tanh\left(kx_i^T x + \theta\right) \tag{2.9}$$

2.2 Artificial Bee Colony Algorithm

In 2005, Karaboga [10] designed the Artificial Bee Colony (ABC) algorithm to effectively tackle unimodal and multi-modal optimization problems. ABC rapidly became one of the latest swarm intelligence techniques. It draws inspiration from the smart behavior of honeybees, with an artificial bee colony comprised of three types of bees: half of the colony comprises working b: employed, onlooker and scout bees. The number of food sources or nectar sources corresponds to the number of working bees, meaning each working bee is linked to one nectar source. The ultimate objective for the entire bee colony is to optimize the nectar yield. The primary responsibility of the employed bees is to search for food sources, essentially the problem solutions. Once they've discovered these sources, they evaluate the amount of nectar or get the solutions, typically expressed as a fitness value. This information is then communicated to the onlooker bees who are stationed within the beehive, waiting to receive and process this data. The onlooker bees decide whether to exploit a nectar source based on the data they receive from the employed bees. They also determine which source to abandon and assign the

employed bee to become a scout bee. The scout bees, in turn, are responsible for discovering new potential food sources. They explore the area around the hive in a random manner. In the ABC algorithm, the problem's solution space is inferred to be D-dimensional, with D representing the number of parameters that need to be optimized. The fitness value of a site chosen at random is as follows:

$$fit_i = \frac{1}{1 + obj.fun_i} \tag{2.10}$$

The number of employed bees and onlooker bees matches the count of food sources guaranteeing a one-to-one ratio between employed bees and food sources. Initially, the positions for both groups are randomly assigned. In each repetition of the ABC algorithm, every employed bee discovers a fresh next food source to its current assignment and evaluates the nectar yield of this newfound source given by:

$$v_{ij} = x_{ij} + \theta \left(x_{ij} - x_{kj} \right) \tag{2.11}$$

where $i = 1, 2, \ldots, \mathrm{SN}, \mathrm{j} = 1, 2, \ldots, \mathrm{D}$ and θ = random number in the interval $[0, 1]$

$$x_i^j = x_{\min}^j + r \left(x_{\max}^j - x_{\min}^j \right) \tag{2.12}$$

where r is a random number in the interval $[0, 1]$. $\mathrm{x}_{\min}^{\mathrm{j}}, \mathrm{x}_{\max}^{\mathrm{j}}$ represent lower and upper borders in the j^{th} dimension respectively, for the problem space.

2.3 Related Work

Researchers introduced several methods to build prediction models based on advanced machine-learning techniques. In [9], a hybrid deep learning framework was introduced to design a joint hybrid deep learning framework that incorporates both one-dimensional CNN and bi-directional LSTM components to extract distinctive features from time series data effectively. They conducted extensive experimental testing using two pollutant datasets demonstrating that this method is robust for PM2.5 forecasting with high accuracy. In [10], combining KNN with particle swarm optimization (PSO) demonstrated good results in air quality classification or prediction. Using an artificial bee colony algorithm combined with machine learning techniques provided a viable alternative to solve extremely complicated, unconstrained continuous optimization problems and prediction nonlinear problems [11, 12]. In [13], the ABC algorithm was employed to optimize functions with multiple variables. It was pitted against other algorithms like

Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Particle Swarm-Inspired Evolutionary Algorithm (PS-EA). In [14], an extended version of ABC proved to be superior, especially for solving constrained optimization problems. For rapid satellite image segmentation, a modified ABC algorithm was utilized. The experimental findings demonstrated that the modified ABC method performed significantly better than the ABC, PSO, and GA methods. In particular, the modified ABC method exhibited a much faster execution time, significantly reducing CPU usage [15]. In [16] ,Particle Swarm Optimization (PSO) within the context of ABC was employed as a cluster routing algorithm to determine the cluster heads and their respective positions concerning the base station. However, ABC's inadequate search capabilities make it a struggle to solve constraint optimization issues effectively. The Artificial Bee Colony Algorithm with Distant Savants (ABCDS), a variation of ABC designed for handling limited optimization issues was suggested in [17] as a solution. The ABCDS system allows learning with savants who are located at specific distances from one another. In [18, 19], an adapted ABC algorithm utilizing a fuzzy multi-objective approach was introduced to address the hyper flow problem.

2.4 Area and Data

Utilizing the three-year hourly dataset (spanning from 2020 to 2022) obtained from the New Cairo Station, situated to the east of Cairo city, serves as the foundation for constructing a robust prediction model aimed at forecasting air quality parameters. To ensure the model's effectiveness, the dataset was meticulously partitioned into two distinct sets: eighty percent of the data was allocated for training the model, and the remaining twenty percent was reserved for testing the model's predictive capabilities. The prediction model focuses on five critical pollutant elements, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), and carbon monoxide. These pollutant elements were selected based on their significant impact on air quality and their relevance to environmental and public health concerns. The utilization of a three-year dataset allows the model to capture seasonal and temporal variations in pollutant concentrations, enabling a comprehensive understanding of air quality dynamics in the New Cairo area. The partitioning of the data into training and testing sets ensures that the model is trained on a representative portion of the dataset and evaluated on unseen data, promoting its generalization and predictive accuracy. This approach contributes to the development of a reliable and effective predictive model for monitoring and managing air

quality in the specified region.

2.5 Proposed Model

Prior to embarking on the model construction process, the dataset, elucidated in Section 2, underwent rigorous and meticulous preprocessing operations to enhance its quality and relevance. A comprehensive array of statistical analysis was meticulously conducted, and the results are meticulously documented in Table 1. These preliminary operations were imperative in deciphering intricate patterns, uncovering hidden insights, and extracting meaningful features from the raw data. The statistical analysis not only laid the groundwork for a more profound understanding of the dataset but also served as a critical precursor to the subsequent modeling endeavors. This proactive approach to data preprocessing underscores our commitment to ensuring the integrity, accuracy, and value of the dataset, setting the stage for the development of a robust and insightful predictive model.

Table 1 sample of statistical analysis

Pollutant	Average (μ)	Standard $Dev(\sigma)$	Missed data
SO2	19.9304	17.9304	103
NO2	57.0103	21.5204	274
CO	7.9216	1.352	216

Data preparation is an important process to complete model building; Weka Machine Learning Workbench was used for this purpose. Weka is a compilation of machine learning algorithms designed for the purpose of data mining. It includes utilities for tasks such as data preprocessing, classifying, regression, clustering, mining association rules, and even data visualization [21]. By Weka all missing values were filled as well as discover and address outliers and extreme values by using unsupervised filters.

The normalization process for all data sets attributes was done by the Equation:

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2.13}$$

Subsequently, the Least Squares Support Vector Machine (LSSVM), fine-tuned through the optimization process facilitated by the Artificial Bee Colony (ABC) algorithm, is employed to meticulously construct a predictive model.

This model is specifically designed to forecast hourly concentration values for pivotal pollutants; namely, sulfur dioxide (SO2), nitrogen dioxide (NO2), and carbon monoxide (CO). The efficacy of the proposed model is rigorously assessed using Root Mean Square Error (RMSE) and Mean Absolute Error (MAS) as robust metrics for evaluating predictive accuracy. The intricate phases involved in the meticulous development of this predictive model are succinctly encapsulated in the visual representation depicted in Fig. 1, providing a comprehensive overview of the methodical and strategic approach undertaken to ensure the model's precision and reliability in predicting pollutant concentrations. This integration of LSSVM, optimized by the adaptive ABC algorithm, and the subsequent evaluation metrics signifies a sophisticated and thoroughly validated methodology, emphasizing our commitment to delivering a predictive model of utmost efficiency and efficacy.

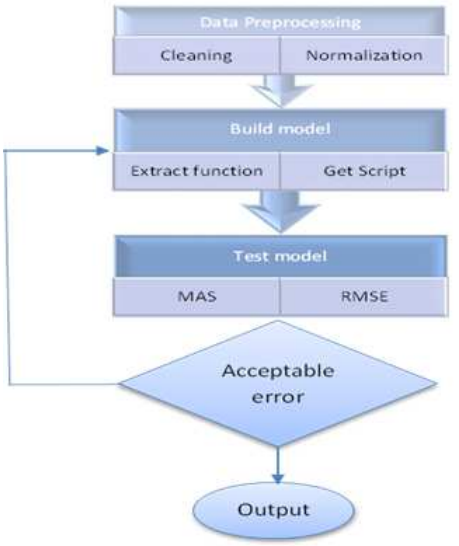

Figure 1: Phases of the proposed model.

The model performance criteria done by calculating RMSE and MAE are given in Table 2:

Table 2 performance evaluations criteria used to Judge

Performance criteria	Symbol	Formula
Root Mean Square Error	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)}$
Mean Absolute Error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert A_i - F_i\rvert$

3 Results

The proposed model tested through unseen data set for the three pollutants SO2, NO2, and CO and the final acceptable results summarized in Table 3.

Table 3 Presented model against MLP performance

	Presented model performance		MLP performance	
Pollutant	MAE	RMSE	MAE	RMSE
SO2	0.5852	0.9431	1.3858	1.8869
NO2	8.3029	11.3914	28.884	17.5273
CO	0.5545	0.8486	0.5851	0.9431

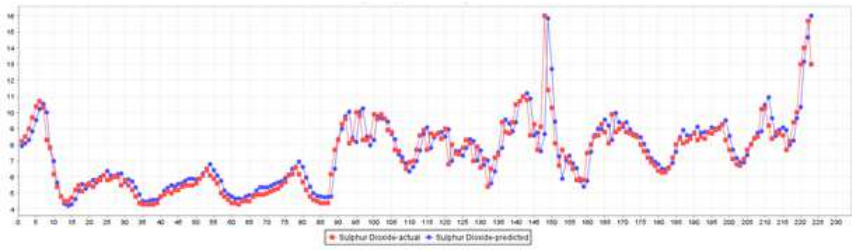

Figure 2: SO2 result.

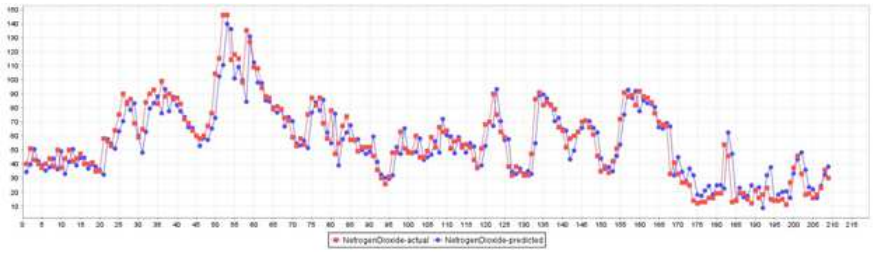

Figure 3: No2 result.

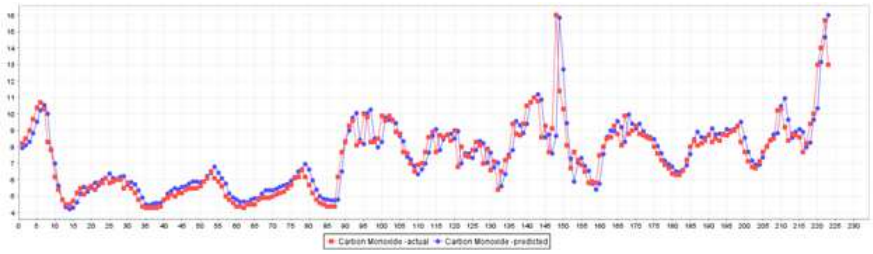

Figure 4: Co result.

The model that was introduced demonstrated significantly superior performance compared with MLP. This superiority became evident when examining the resulting Mean Absolute Error (MAS) and Root Mean Square Error (RMSE) values for both methods. MLPs tend to be less proficient when applied to tasks such as analyzing time series data and predicting nonlinear variables because they were unable to grasp the sequential relationships within the data. Figures 2, 3, and 4 display the output results, illustrating the predicted values achieved by the presented model for SO2, NO2, and CO, respectively.

Moreover, Figures 2, 3, and 4 illustrate the hourly forecasted output for SO2, NO2, and CO. On the Y-axis, the concentration of pollutants measured in micrograms per cubic meter ($\mu g/m^3$), while the X- axis represents the time steps, when the model applied on the test data set. The red line depicts the data, and the blue line represents the predicted hourly values. These line graphs clearly depict the model's impressive accuracy in predicting pollutant levels.

4 Conclusions

Pollutant levels can vary significantly both spatially and temporally. Atmospheric conditions, localized sources of pollution, and seasonal variations can lead to substantial fluctuations in pollutant concentrations. Accurately capturing these variations is a challenge, requiring unusual techniques to predict. ABC uses a stochastic search process where employed onlooker bees investigate the search space. This adaptability helps the algorithm overcome challenges in different types of optimization problems. Hybrid LSSVM-ABC models for pollutants prediction is presented in this work to predict three pollutants (SO2, NO2, and CO). The presented model demonstrated high performance and efficiency compared with MLP.

Acknowledgment. The authors gratefully acknowledge the approval and the support of this research study by grant no SCAT-2022-11-1529 from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A.

References

[1] A. Azam, B. Riahi-Zanjani, M. Balali-Mood", Effects of air pollution on human health and practical measures for prevention in Iran,", J Res. Med. Sci., **1,** (2016), 21–65.
doi: 10.4103/1735-1995.189646. PMID: 27904610; PMCID: PMC5122104.

[2] Luisa T. Molina, Tong Zhu, Wei Wan, Bhola Ram Gurjar. "Impacts of Megacities on Air Quality: Challenges and Opportunities." Environmental Science, (2020).

[3] X. Wang, M. Meng, "A Hybrid Neural Network and ARIMA Model for Energy Consumption Forecasting", Journal of computers, **7,** no. 5, (2012), 1184–1190.

[4] Gangavarapu Sailasya, Gorli L Aruna Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms", International Journal of Advanced Computer Science and Applications, **12,** no. 6, (2021).

[5] S. A. Asklany, Khaled Elhelow, M . Abd El Wahab, "On using Adaptive Hybrid Intelligent Systems in PM10 Prediction", International Journal of Soft Computing and Engineering, **6,** (2016) 54–59.

[6] S. Fan, D. Hao, Y. Feng, K. Xia, W. Yang, "A Hybrid Model for Air Quality Prediction Based on Data Decomposition", Information, **12,** (2021).

[7] Jinghua Li, Yongsheng Lei, Shuhui Yang, "Mid-long term load forecasting model based on support vector machine optimized by improved sparrow search algorithm", Energy Reports, **8,** (2022), 491–497.

[8] J. Zeng, Z-H Tan, T. Matsunaga, T. Shirai, Generalization of Parameter Selection of SVM and LS-SVM for Regression. Machine Learning and Knowledge Extraction, **1,** (2019), 745–755.

[9] S. Du, T. Li, Y. Yang, S. J. Horng, "Deep Air Quality Forecasting Using Hybrid Deep Learning Framework," IEEE Transactions on Knowledge and Data Engineering, **33,** no. 6, (2021), 2412–2424.

[10] S. Yahdin, A. Desiani, S. P. Andini, D. Cahyawati, M. Arhami, N. Eliyati, "Combination of KNN and Particle Swarm Optimization on air quality prediction", Barekeng: J. Il. Mat. & Ter., **16,** no. 1, (2022), 7–14.

[11] A. Kumar Bandrana, G. Kabra, E. K. Mussada, M. K. Dash, P. S. Rana, Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention. Neural Comput & Applic, **31,** (2019), 877–890.

[12] Min-Rong Chen, Jun-Han Chen, Guo-Qiang Zeng, Kang-Di Lu, Xin-Fa Jiang, An improved artificial bee colony algorithm combined with extremal optimization and Boltzmann Selection probability, Swarm and Evolutionary Computation, **49,** (2019), 158–177.

[13] D. Karaboga, B. Basturk, "A Powerful And Efficient Algorithm For Numerical Function Optimization: Artificial Bee Colony Algorithm", Journal of Global Optimization, **39,** no. 3, (2007), 459–471.

[14] D. Karaboga, B. Basturk, On The Performance Of Artificial Bee Colony (ABC) Algorithm, Applied SoftComputing journal, **8,** no. 1, (2008), 687–697.

[15] A. K. Bhandari, A. Kumar, G. K. Singh, "Modified Artificial Bee Colony Based Computationally Efficient Multilevel Thresholding for Satellite Image Segmentation Using Kapur's, Otsu and Tsallis Functions", Pergamon Press Inc., **42,** no. 3, 2015.

[16] L. Sixu, W. Muqing, Z. Min, "Particle swarm optimization and artificial bee colony algorithm for clustering and mobile based software-defined wireless sensor networks", Wireless Network, **28,** (2022), 1671–1688.

[17] Gürcan Yavuz, Burhanettin Durmuş, Doğan Aydın, Artificial Bee Colony Algorithm with Distant Savants for constrained optimization, Applied Soft Computing, **116,** (2022).

[18] M. Li, G. G. Wang, H. Yu," Sorting-Based Discrete Artificial Bee Colony Algorithm for Solving Fuzzy Hybrid Flow Shop Green Scheduling Problem", Mathematics, **9,** (2021), 2250.

[19] J. Suykens, J. Vandewalle, "Least squares support vector machine classifiers", Neural Processing Letters, **9,** no. 37, (1999), 293–300.