Structured query language injection detection with natural language processing techniques optimized by metaheuristics

Aleksandar Jokic[1], Nikola Jovic[1], Vuk Gajic[1], Marina Svicevic[2] Milos Pavkovic[1], and Aleksandar Petrovic[1]*

[1] Singidunum University, Faculty of Informatics and Computing, Danijelova 32, 11000 Belgrade, Serbia
aleksandar.jokic.22@singimail.rs, nikola.jovic.23@singimail.rs, vgajic@singdinum.ac.rs, mpavkovic@singdinum.ac.rs, aleksandar.petrovic@singdinum.ac.rs*

[2] University of Kragujevac, Faculty of Science, Radoja Domanovića 12, 34000 Kragujevac, Serbia
marina.svicevic@pmf.kg.ac.rs

Abstract. This research focuses on the detection of Structured Query Language (SQL) injection intrusion detection. This problem has gained significance due to the widespread use of SQL in different systems, as well as for the numerous versions of attacks that are performable by using this technique. This work aims to propose a robust solution for the detection of such attacks by applying artificial intelligence (AI). The data is preprocessed by a Bidirectional Encoder Representations from Transformers (BERT), while the predictions are made by the Extreme Gradient Boosting (XGBoost) algorithm. The XGBoost is a powerful predictor if optimized correctly. Hyperparameters are optimized by an improved version of the Crayfish Optimization Algorithm (COA) hybridized with the Genetic Algorithm (GA). The proposed solution is tested against high-performing metaheuristics in which it achieved favorable performance.

Keywords: sql injection · swarm intelligence · coa · natural language processing · BERT · XGBoost

1 Introduction

Data querying, manipulation, definition, control, and administration define the basic functions of the (SQL). Such operations are performed in communication with relational databases. The data stored is structured and can have connections between its entities that provide additional information on the data. The creators of SQL are Donald D. Chamberlin and Raymond F. Boyce [18]. The simplicity of the language and its efficiency are relevant despite the fact that it was released in the 1970s. Different standards have appeared since, including ANSI and ISO since 1986 and 1987 [24]. The popularity of SQL and its widespread

N. Bacanin and H. Shaker (eds.), *Proceedings of the 2nd International Conference on Innovation in Information Technology and Business (ICIITB 2024)*,
Advances in Computer Science Research 113,
https://doi.org/10.2991/978-94-6463-482-2_11

use unfortunately make it an interesting target for attackers. Another reason for why it is often chosen as a means for intrusion is its backward compatibility.

Natural language processing (NLP) is one of many fields that can benefit from the application of AI. The use of AI with SQL queries can have many benefits of which the most important is security. Automation is another important reason for AI integration into relational database management systems (RDBMS), although this process is not without challenges [41].

To begin with, the hyperparameter optimization problem is paramount to the success of hybrid solutions. Deep learning is widely applied but it requires tuning of its parameters and architecture if the highest performance is to be achieved. The process of optimization is performed through the use of advanced metaheuristic optimizers which undergo extensive testing.

The application of metaheuristics inspired by nature leads to significant improvements in the optimization of hyperparameters. The genetic algorithm (GA) [47] represents a prime example of such solutions, followed by the particle swarm optimization [70], and the red fox algorithm [48].

Metaheuristic solutions are efficient in searching large areas of solutions and provide optimal values of hyperparameters. This problem is considered to have NP-hard complexity. This needs to be achieved while not disrupting the existing SQL protocols and standards. For safe integration with the SQL systems, the application of AI must be efficient and effective.

The main predictor of the proposed framework for SQL injection detection is the XGBoost algorithm, which is optimized by the proposed genetically inspired crayfish optimization algorithm (GICOA), while BERT is responsible for data processing.

The highest influential contributions of this work are provided in the list:

- Potential exploration of optimized NLP models for SQL injection detection.
- Testing of BERT for SQL injection detection use case.
- The use of advanced and modified metaheuristic optimizers for performance optimization.

The structure of the paper is: Section 2 gives the fundamentalas of the technologies applied in this work, Section 3 presents the basic version of the applied metaheuristics followed by the modifications applied to it, Section 5 holds the information necessary for experiment recreation, while the Section 6 provides the outcomes, and the last Section 7 concludes the paper.

2 Related works

In the work of Crespo et al. [22] flow data has been detected and collected into two datasets by lightweight protocols. The data was collected from the most popular database engines. The results indicate a false positive rate below 0.07% with the detection rate over 97%.

Lu et al. [42] propose synBERT which is a detection model based on semantic learning for SQL injection detection. The model is capable of mapping SQL syntax tree structures from learned representations. The model indicated excellent generalization performance and has been tested on multiple datasets. The authors collected a dataset with a wide variety of SQL injection types, which is paramount due to their large number.

Alarfaj et al. [1] tackle the problem of detection mechanism for SQL injection through use of a probabilistic neural network. The neural network has been optimized by the Bat algorithm, and tokenization and regular expression were used for feature extraction. The data used in the research was collected from network traffic. The reported metrics' results are 99.19% for accuracy, 0.995% for precision, 0.981% for recall, and 0.928% for F-measure over 10-fold cross-validation.

The use of AI in cybersecurity is on the rise, especially when considering the rapid advancing Internet of Things (IoT) technologies. The important sector that tends to use more and more smart devices is healthcare. Savanovic et al. [60] provide a solution for intrusion detection with the goal of improving the rate in which this field advances. The authors proposed an improved swarm metaheuristic based on the firefly algorithm (FA). The main predictor that was used in this research is the XGBoost which is optimized by the said improved solution for optimal performance.

On the other hand, Zivkovic et al. [74] explored the XGBoost solution optimized by the Since Cosine Algorithm (SCA) for the same problem, while in the work of Petrovic et al. [49] Multi-verse Optimizer Algorithm (MOA) is applied with improvements to it.

Bacanin et al. [15] explore a different type of NLP use with AI. In their work spam email are classified by yet another XGBoost approach. The main predictor is optimized by the Social Network Search (SNS) algorithm and tested on spam benchmark datasets against other high-performing metaheuristics.

The hyperparameter optimization is crucial for the performance of the machine learning models [6]. Consequentially, for every different use case, a new model needs to be picked and optimized. The No Free Lunch (NFL) theorem states that there is not a universal solution that can solve all problems [71].

Significant use cases of metaheuristic optimizers applications are reported in the following text: plant classification[14, 83], credit card frauds identification [30, 51, 50] wireless sensor networks optimization [73, 4], cloud computing problems [56, 9], global optimization problems [17, 11], automotive traffic predictions [54, 26], medicine [40, 72, 39, 79, 33, 81, 78, 46], economic problems [34, 38, 65, 28, 55, 53], environmental monitoring and pollutants tracking [44, 32, 8, 43, 23], spam emails [77, 57], predicting green energy production [62, 5, 37], feature selection [80, 16, 31, 76, 7], marine vessel classification and trajectory prediction [52, 67], defect identification in software testing [84], enhancing the audit opinion [66], computer security, phishing and intrusion detection [75, 2, 60, 36, 59, 15, 35, 12], and general optimization of machine learning models [10, 61, 13, 45, 58, 82, 3, 27, 69], and others [64, 63, 68].

2.1 BERT

The BERT technical solution was introduced in 2018 [25] and its basis is the attention mechanism for text and sentence symbolism interpretation. The solution was created by a research team from Google, and its success testifies to the number of NLP applications that apply it for purposes like speech recognition, translation, and many more.

The transformers are the basis of BERT's architecture that thanks to the attention mechanism can shift its focus between the input data segments during processing. The technical solution aids the understanding of the sentence as it provides the meaning of the words. The attention mechanism is applied in parallelization allowing for processing of different sentence parts at the same time. This particularly increases the efficiency and speed of data processing.

BERT is capable of processing natural language bidirectionally, which makes it possible to grasp the full context of a word by analyzing what comes before as well as after. This technique is paramount for understanding the deeper meaning of the language itself, as the previous model only analyzed text in one direction.

The model is trained with the use of the "masked language model" (MLM). The concept of the technique is such that a random word is hidden from the model while it has to predict it only from the content provided. By doing so, the model is capable of language patterns and structure understanding.

By preparing a BERT model, it can be efficiently applied for special tasks with its support for transfer learning if trained on a large dataset. Text classification, summarization, and querying are some of the most notable applications of BERT in NLP. Consequentially, BERT is an exemplary choice for many different tasks of such type. Its role in intelligent systems is important due to its ability to understand and interpret human language.

2.2 XGBoost

The XGBoost is distinguished from other machine learning methods thanks to its high performance [20, 19]. Nonetheless, this performance is achieved only if proper hyperparameters are selected, which in terms of the problem's nature belongs to an extremely difficult group of problems recognized as NP-hard. XGBoost model consists of many weaker predictors in combination for making an accurate model. Significant improvements in performance are achievable when gradient boosting and regularization are applied. The model makes predictions from learned patterns with an accent on complex input and target dependency management. The XGBoost's objective functions is provided in Equation 1.

$$1.\text{obj}(\Theta) = L(\theta) + \Omega(\Theta), \tag{1}$$

in which the loss function and regularization are combined. Θ represents the hyperparameter set represented, the loss function as $L(\Theta)$, while the complexity of the model is managed by regularization term $\Omega(\Theta)$.

Mean square error (MSE) is used for loss function application as shown in Equation 2.

$$\mathrm{L}(\Theta) = \sum_i (y_i - \hat{y_i})^2, \tag{2}$$

in which the y_i indicates the predicted value, while the $\hat{y_i}$ predicted target variable's value for each iteration i.

The differentiation process of actual and predicted values is provided in Equation 3. The classification's results increase with the minimization of the overall loss function.

$$L(\Theta) = \sum_i [\, y_i \ln\, (1 + e^{-\hat{y}_i}) + (1 - y_i) \ln\, (1 + e^{\hat{y}_i})]. \tag{3}$$

3 Methods

The novel crayfish optimization algorithm (COA) [29] is exploited in this research as the optimizer for the XGBoost algorithm. In this Section, the original COA is described along with equations that are used to model crayfish behavior. Afterward, the modifications to the original COA are described and the proposed GICOA is introduced.

3.1 Original COA

Metaheuristic-based algorithms consist of a pair or pairs of phases, that consist of exploration and exploitation phases, and such is the case with the COA. The crayfish avoid the waters where the temperature exceeds 30 degrees Celsius and this behavior is used to model the exploration phase. For the exploitation phase, the competitive nature of the crayfish is modeled. This behavior occurs when two specimens in proximity compete over a cave. After the temperature returns below 30 degrees, the crayfish leaves the cave and begins foraging.

The cave in the exploration phase is described be the following equation:

$$X_{\text{shade}} = \frac{X_G + X_L}{2} \tag{4}$$

in which the optimal location over rounds is X_G, and the current solutions set's ideal location is X_L.

When the value $rand$ is less than 0.5 it is simulated that there are no crayfish to fight against and the unit joins a cave, modeled by the equation:

$$X_{i,j}^{t+1} = X_{i,j}^{t} + C_2 \times \text{rand} \times (X_{\text{shade}} - X_{i,j}^{t}) \tag{5}$$

where the current round of execution is denoted as t, the next generation's following round number as $t + 1$, and the C_2 is calculated by the next equation:

$$C_2 = 2 - \frac{t}{T} \tag{6}$$

in which the maximum run number is T.

In the case where $rand$ is above 0.5, the next equation describes the competing behavior:

$$X_{i,j}^{t+1} = X_{i,j}^{t} - X_{z,j}^{t} + X_{\text{shade}} \quad (7)$$

where an arbitrary crayfish is represented by z, which is modeled by the following equation:

$$z = \text{round}(\text{rand} \times (N - 1)) + 1 \quad (8)$$

When the water temperature is ideal and drops bellow 30 degrees, the crayfish enter the feeding stage. $X_{food} = X_G$ denotes the position of the food, and the process of foraging is modeled as follows:

$$X_{(13),\, t+1_{i,j}} = X_{t_{i,j}} + X_{\text{food}} \times p \times (\cos(2 \times \pi \times \text{rand}) - \sin(2 \times \pi \times \text{rand})) \quad (9)$$

In some cases the food is too large to be consumed as is and the crayfish chops it down. The condition $Q \leq \frac{C3+1}{2}$ is used to model the boundary for food size, which if true the crayfish will have to break down its food as seen in:

$$X_{\text{food}} = \exp\left(\frac{-1}{Q}\right) \times X_{\text{food}} \quad (10)$$

in which the food size is denoted by Q and described as:

$$Q = C3 \times \text{rand} \times \left(\frac{\text{fitness}_i}{\text{fitness}_{\text{food}}}\right) \quad (11)$$

in which the largest food unit is C_3 and it has a constant value 3, the i-th's crayifhs fitness is given as $fitness_i$, and teh $fitness_{food}$ shows the food source position's fitness value.

When the condition for food size, $Q \leq \frac{C3+1}{2}$, is not satisfied the behavior of eating is modeled as:

$$X_{i,j}^{t+1} = (X_{i,j}^{t} - X_{\text{food}}) \times p + p \times \text{rand} \times X_{i,j}^{t} \quad (12)$$

4 Genetically inspired COA

Genetically inspired COA (GICOA) is the name the authors give to the proposed improved solution. While the original COA has solid performance in exploration as well as exploitation, the tradeoff between them can be improved. This is attempted with mutation and crossover genetic properties. For the first $max_iter/2$ the two worst solutions are replaced without additional evaluation with the crossover being performed between the best solution and the quasi reflexive learning (QRL) opposite of the second best solution with the probability for crossover $p_c = 0.05$. After this two offspring are created with a mutation probability of $p_m = 0.05$. The general mutation for every parameter is performed in the case of $rnd > 0.5$ where the lower bound of that parameter is divided by 10 and subtracted from it. Otherwise, it is added to it. In the second half of all iterations, the two worst solutions are replaced in the same manner except the crossover is performed between the two best solutions.

5 Experimental setup

A public dataset was applied for this research https://www.kaggle.com/datasets/sajid576/sql-injection-dataset, and it is formed of raw SQL queries of two groups. The 0 group represents non-malicious data, while the class 1 represents malicious data. 70% of the dataset is used for the training counterpart, and the rest is used for testing.

Due to the imbalances in the dataset, the authors use the Cohen's κ as an indicator function which is to be maximized [21]. The following equation describes this metric:

$$\kappa = \frac{c_o - c_e}{1 - c_e} = 1 - \frac{1 - c_o}{1 - c_e} \tag{13}$$

where the c_o and c_e represent the observed and expected vectors respectively. The results of using this metric indicate robust prediction in comparison to only using the accuracy.

6 Outcomes

Tables 1 and 2 provide the results for the objective and indicator function test results, respectively. The best results are indicated with bold font in all tables. In both mentioned Tables, for the best metric the proposed solution, XG-GICOA achieved the best result. It was outperformed by the XG-SCHO across other metrics. However, the proposed solution had better results across all categories than the original COA.

Table 1. SQL injection detection objective function outcomes for each optimizer.

Method	Best	Worst	Mean	Median	Std	Var
XG-GICOA	**.992710**	.988886	.990314	.989931	.001167	1.36E-06
XG-COA	.990977	.989585	.990245	.990278	.000548	3.01E-07
XG-PSO	.990977	.989233	.990244	.990453	.000592	3.50E-07
XG-ABC	.990629	.988890	.990106	.990278	.000473	2.23E-07
XG-FA	.992368	.989929	.990660	.990275	.000830	6.89E-07
XG-SCHO	.992016	**.990277**	**.991218**	**.991321**	**.000441**	**1.95E-07**
XG-EHO	.991669	.989583	.990175	.989932	.000677	4.59E-07
XG-RSA	.991669	.989931	.990662	.990455	.000590	3.48E-07

7 Conclusion

SQL injection is a problem due to its widespread use and the vulnerabilities that it creates. Traditional security measures were not able to detect malicious intent in a similar manner that AI does. With the use of a framework for advanced

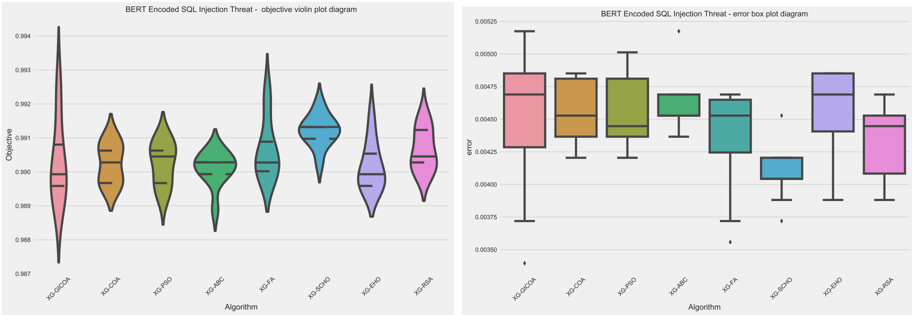

Fig. 1. Objective and indicator function outcomes distributions for SQL injection detection.

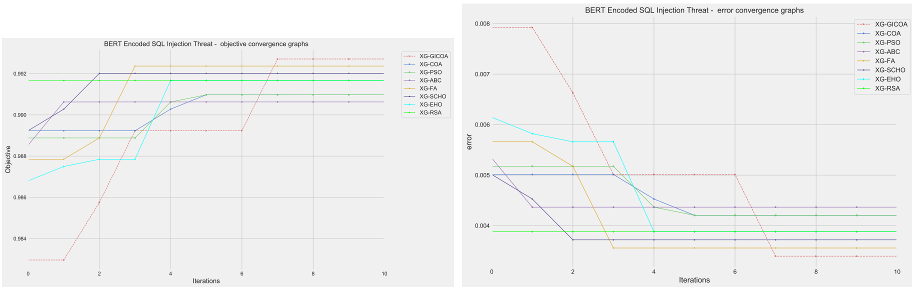

Fig. 2. Objective and indicator function convergences for SQL injection detection.

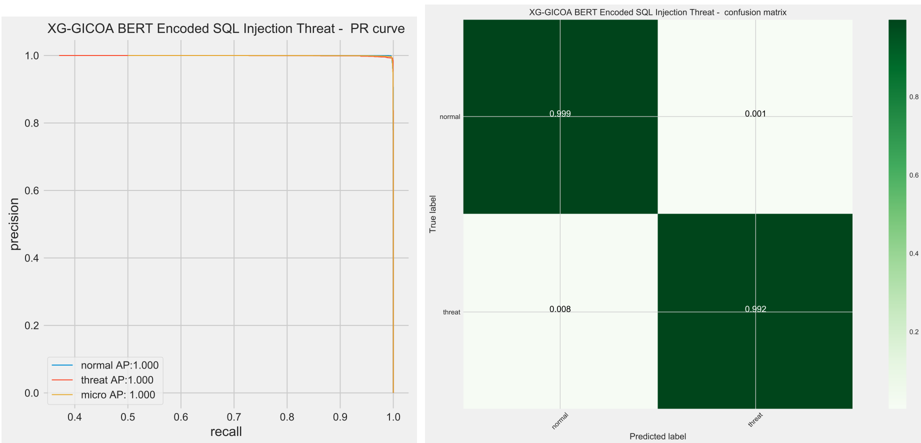

Fig. 3. Best constructed model PR plot and confusion matrix.

Table 2. SQL injection detection indicator function outcomes for each optimizer.

Method	Best	Worst	Mean	Median	Std	Var
XG-GICOA	**.003396**	.005175	.004512	.004690	.000543	2.95E-07
XG-COA	.004204	.004851	.004544	.004528	.000255	6.51E-08
XG-PSO	.004204	.005013	.004544	.004447	.000275	7.56E-08
XG-ABC	.004366	.005175	.004609	.004528	.000220	4.84E-08
XG-FA	.003558	.004690	.004350	.004528	.000386	1.49E-07
XG-SCHO	.003719	**.004528**	**.004091**	**.004043**	**.000205**	**4.21E-08**
XG-EHO	.003881	.004851	.004576	.004690	.000316	9.96E-08
XG-RSA	.003881	.004690	.004350	.004447	.000275	7.56E-08

Table 3. SQL injection detection detailed metrics for each optimizer best constructed model.

Method	metric	injection	non-injection	accuracy	macro avg.	weighted avg.
XG-GICOA	precision	.995394	.998682	.996604	.997038	.996612
	recall	.999229	.992143	.996604	.995686	.996604
	f1-score	.997308	.995402	.996604	.996355	.996602
XG-COA	precision	.995134	.996928	.995796	.996031	.995799
	recall	.998202	.991707	.995796	.994954	.995796
	f1-score	.996666	.994311	.995796	.995488	.995793
XG-PSO	precision	.995134	.996928	.995796	.996031	.995799
	recall	.998202	.991707	.995796	.994954	.995796
	f1-score	.996666	.994311	.995796	.995488	.995793
XG-ABC	precision	.994880	.996927	.995634	.995903	.995638
	recall	.998202	.991270	.995634	.994736	.995634
	f1-score	.996538	.994091	.995634	.995314	.995631
XG-FA	precision	.996155	.996934	.996442	.996544	.996443
	recall	.998202	.993453	.996442	.995827	.996442
	f1-score	.997177	.995190	.996442	.996184	.996441
XG-SCHO	precision	.995138	.998243	.996281	.996690	.996288
	recall	.998973	.991707	.996281	.995340	.996281
	f1-score	.997052	.994964	.996281	.996008	.996278
XG-EHO	precision	.995137	.997804	.996119	.996471	.996125
	recall	.998716	.991707	.996119	.995211	.996119
	f1-score	.996923	.994746	.996119	.995835	.996117
XG-RSA	precision	.995137	.997804	.996119	.996471	.996125
	recall	.998716	.991707	.996119	.995211	.996119
	f1-score	.996923	.994746	.996119	.995835	.996117
	support	3893	2291			

Table 4. Parameter selections made by each optimizer for the respective best performing models.

Method	Learning Rate	Min Child W.	Subsample	Col by Tree	Max depth	Gamma
XG-GICOA	.898699	3.493451	.617036	.875158	7	.050642
XG-COA	.900000	1.000000	1.000000	.618234	5	.155831
XG-PSO	.837173	5.759070	1.000000	.594179	10	.716683
XG-ABC	.900000	5.258180	.579786	.585100	9	.716699
XG-FA	.900000	9.523153	.619419	1.000000	8	.800000
XG-SCHO	.900000	1.589997	1.000000	.836143	8	.107443
XG-EHO	.900000	5.130561	.669494	.386246	10	.800000
XG-RSA	.900000	10.000000	.872518	.945474	7	.600513

pattern recognition, it is possible to detect such malicious intent with high precision. This work provides a novel hybrid metaheuristic solution rooted in swarm intelligence. The original algorithm used is the COA and it has been modified with the QRL and the principles from the GA. The proposed solution has been tested against other high-performing solutions and it has shown admirable performance. It is important to note that the original COA did not provide better performance than the proposed GICOA. The results indicate that there is still space for improvement and future work will focus on other possible combinations with the COA as well as other solutions.

References

1. Alarfaj, F.K., Khan, N.A.: Enhancing the performance of sql injection attack detection through probabilistic neural networks. Applied Sciences **13**(7), 4365 (2023)
2. AlHosni, N., Jovanovic, L., Antonijevic, M., Bukumira, M., Zivkovic, M., Strumberger, I., Mani, J.P., Bacanin, N.: The xgboost model for network intrusion detection boosted by enhanced sine cosine algorithm. In: International Conference on Image Processing and Capsule Networks. pp. 213–228. Springer (2022)
3. Bacanin, N., Antonijevic, M., Vukobrat, N., Bezdan, T., Zivkovic, M.: Enhanced seagull optimization algorithm adapted for artificial neural network training. In: ICT with Intelligent Applications: Proceedings of ICTIS 2021, Volume 1. pp. 753–761. Springer (2022)
4. Bacanin, N., Arnaut, U., Zivkovic, M., Bezdan, T., Rashid, T.A.: Energy efficient clustering in wireless sensor networks by opposition-based initialization bat algorithm. In: Computer Networks and Inventive Communication Technologies: Proceedings of Fourth ICCNCT 2021. pp. 1–16. Springer (2022)
5. Bacanin, N., Jovanovic, L., Zivkovic, M., Kandasamy, V., Antonijevic, M., Deveci, M., Strumberger, I.: Multivariate energy forecasting via metaheuristic tuned long-short term memory and gated recurrent unit neural networks. Information Sciences **642**, 119122 (2023)
6. Bacanin, N., Petrovic, A., Antonijevic, M., Zivkovic, M., Sarac, M., Tuba, E., Strumberger, I.: Intrusion detection by xgboost model tuned by improved social network search algorithm. In: International Conference on Modelling and Development of Intelligent Systems. pp. 104–121. Springer (2022)
7. Bacanin, N., Petrovic, A., Zivkovic, M., Bezdan, T., Chhabra, A.: Enhanced salp swarm algorithm for feature selection. In: International conference on intelligent and fuzzy systems. pp. 483–491. Springer International Publishing Cham (2021)
8. Bacanin, N., Sarac, M., Budimirovic, N., Zivkovic, M., AlZubi, A.A., Bashir, A.K.: Smart wireless health care system using graph lstm pollution prediction and dragonfly node localization. Sustainable Computing: Informatics and Systems **35**, 100711 (2022)
9. Bacanin, N., Simic, V., Zivkovic, M., Alrasheedi, M., Petrovic, A.: Cloud computing load prediction by decomposition reinforced attention long short-term memory network optimized by modified particle swarm optimization algorithm. Annals of Operations Research pp. 1–34 (2023)
10. Bacanin, N., Stoean, C., Zivkovic, M., Jovanovic, D., Antonijevic, M., Mladenovic, D.: Multi-swarm algorithm for extreme learning machine optimization. Sensors **22**(11), 4204 (2022)
11. Bacanin, N., Stoean, R., Zivkovic, M., Petrovic, A., Rashid, T.A., Bezdan, T.: Performance of a novel chaotic firefly algorithm with enhanced exploration for tackling global optimization problems: Application for dropout regularization. Mathematics **9**(21), 2705 (2021)
12. Bacanin, N., Zivkovic, M., Hajdarevic, Z., Petrovic, A., Budimirovic, N., Antonijevic, M., Strumberger, I.: Detection of botnet using extreme learning machine tuned by enhanced sine cosine algorithm. In: International Conference on Advances and Applications of Artificial Intelligence and Machine Learning. pp. 125–137. Springer (2022)
13. Bacanin, N., Zivkovic, M., Salb, M., Strumberger, I., Chhabra, A.: Convolutional neural networks hyperparameters optimization using sine cosine algorithm. In: Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021. pp. 863–878. Springer (2022)

14. Bacanin, N., Zivkovic, M., Sarac, M., Petrovic, A., Strumberger, I., Antonijevic, M., Petrovic, A., Venkatachalam, K.: A novel multiswarm firefly algorithm: An application for plant classification. In: International Conference on Intelligent and Fuzzy Systems. pp. 1007–1016. Springer (2022)
15. Bacanin, N., Zivkovic, M., Stoean, C., Antonijevic, M., Janicijevic, S., Sarac, M., Strumberger, I.: Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. Mathematics **10**(22), 4173 (2022)
16. Bezdan, T., Cvetnic, D., Gajic, L., Zivkovic, M., Strumberger, I., Bacanin, N.: Feature selection by firefly algorithm with improved initialization strategy. In: 7th conference on the engineering of computer based systems. pp. 1–8 (2021)
17. Bezdan, T., Petrovic, A., Zivkovic, M., Strumberger, I., Devi, V.K., Bacanin, N.: Current best opposition-based learning salp swarm algorithm for global numerical optimization. In: 2021 Zooming Innovation in Consumer Technologies Conference (ZINC). pp. 5–10. IEEE (2021)
18. Boyce, R.F., Chamberlin, D.D.: Using a structured English query language as a data definition facility. IBM Thomas J. Watson Research Division (1973)
19. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
20. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al.: Xgboost: extreme gradient boosting. R package version 0.4-2 **1**(4), 1–4 (2015)
21. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
22. Crespo-Martínez, I.S., Campazas-Vega, A., Guerrero-Higueras, Á.M., Riego-DelCastillo, V., Álvarez-Aparicio, C., Fernández-Llamas, C.: Sql injection attack detection in network flow data. Computers & Security **127**, 103093 (2023)
23. Damaševičius, R., Jovanovic, L., Petrovic, A., Zivkovic, M., Bacanin, N., Jovanovic, D., Antonijevic, M.: Decomposition aided attention-based recurrent neural networks for multistep ahead time-series forecasting of renewable power generation. PeerJ Computer Science **10** (2024)
24. Date, C.J.: A Guide to the SQL Standard. Addison-Wesley Longman Publishing Co., Inc. (1989)
25. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
26. Dogandzic, T., Petrovic, A., Jovanovic, L., Bacanin, N., Jovanovic, A., Zivkovic, M.: Speeding classification by a deep learning audio analysis system optimized by the reptile search algorithm. In: International Joint Conference on Advances in Computational Intelligence. pp. 73–88. Springer (2022)
27. Gajic, L., Cvetnic, D., Zivkovic, M., Bezdan, T., Bacanin, N., Milosevic, S.: Multi-layer perceptron training using hybridized bat algorithm. In: Computational Vision and Bio-Inspired Computing: ICCVBIC 2020. pp. 689–705. Springer (2021)
28. Golubovic, S., Petrovic, A., Bozovic, A., Antonijevic, M., Zivkovic, M., Bacanin, N.: Gold price forecast using variational mode decomposition-aided long short-term model tuned by modified whale optimization algorithm. In: International Conference on Data Intelligence and Cognitive Informatics. pp. 69–83. Springer (2023)
29. Jia, H., Rao, H., Wen, C., Mirjalili, S.: Crayfish optimization algorithm. Artificial Intelligence Review **56**(Suppl 2), 1919–1979 (2023)

30. Jovanovic, D., Antonijevic, M., Stankovic, M., Zivkovic, M., Tanaskovic, M., Bacanin, N.: Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. Mathematics **10**(13), 2272 (2022)
31. Jovanovic, D., Marjanovic, M., Antonijevic, M., Zivkovic, M., Budimirovic, N., Bacanin, N.: Feature selection by improved sand cat swarm optimizer for intrusion detection. In: 2022 International Conference on Artificial Intelligence in Everything (AIE). pp. 685–690. IEEE (2022)
32. Jovanovic, G., Perisic, M., Bacanin, N., Zivkovic, M., Stanisic, S., Strumberger, I., Alimpic, F., Stojic, A.: Potential of coupling metaheuristics-optimized-xgboost and shap in revealing pahs environmental fate. Toxics **11**(4), 394 (2023)
33. Jovanovic, L., Bacanin, N., Zivkovic, M., Antonijevic, M., Petrovic, A., Zivkovic, T.: Anomaly detection in ecg using recurrent networks optimized by modified metaheuristic algorithm. In: 2023 31st Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2023)
34. Jovanovic, L., Jovanovic, D., Bacanin, N., Jovancai Stakic, A., Antonijevic, M., Magd, H., Thirumalaisamy, R., Zivkovic, M.: Multi-step crude oil price prediction based on lstm approach tuned by salp swarm algorithm with disputation operator. Sustainability **14**(21), 14616 (2022)
35. Jovanovic, L., Jovanovic, D., Antonijevic, M., Nikolic, B., Bacanin, N., Zivkovic, M., Strumberger, I.: Improving phishing website detection using a hybrid two-level framework for feature selection and xgboost tuning. Journal of Web Engineering **22**(3), 543–574 (2023)
36. Jovanovic, L., Jovanovic, D., Antonijevic, M., Zivkovic, M., Budimirovic, N., Strumberger, I., Bacanin, N.: The xgboost tuning by improved firefly algorithm for network intrusion detection. In: 2022 24th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). pp. 268–275. IEEE (2022)
37. Jovanovic, L., Kljajic, M., Petrovic, A., Mizdrakovic, V., Zivkovic, M., Bacanin, N.: Modified teaching-learning-based algorithm tuned long short-term memory for household energy consumption forecasting. In: International Conference on Worldwide Computing and Its Applications. pp. 347–362. Springer Nature Singapore (2023)
38. Jovanovic, L., Milutinovic, N., Gajevic, M., Krstovic, J., Rashid, T.A., Petrovic, A.: Sine cosine algorithm for simple recurrent neural network tuning for stock market prediction. In: 2022 30th Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2022)
39. Jovanovic, L., Petrovic, A., Zivkovic, T., Antonijevic, M., Bacanin, N., Zivkovic, M.: Exploring the potential of generative adversarial networks for synthetic medical data generation. In: 2023 31st Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2023)
40. Jovanovic, L., Zivkovic, M., Antonijevic, M., Jovanovic, D., Ivanovic, M., Jassim, H.S.: An emperor penguin optimizer application for medical diagnostics. In: 2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC). pp. 191–196. IEEE (2022)
41. Li, G., Zhou, X., Cao, L.: Ai meets database: Ai4db and db4ai. In: Proceedings of the 2021 International Conference on Management of Data. pp. 2859–2866 (2021)
42. Lu, D., Fei, J., Liu, L.: A semantic learning-based sql injection attack detection technology. Electronics **12**(6), 1344 (2023)
43. Mani, J., Shaker, H., Jovanovic, L., et al.: Sunspot occurrence forecasting with metaheuristic optimized recurrent neural networks. Theoretical and Applied Computational Intelligence **1**(1), 15–26 (2023)

44. Markovic, I., Krzanovic, J., Jovanovic, L., Toskovic, A., Bacanin, N., Petrovic, A., Zivkovic, M.: Flood prediction based on recurrent neural network time series classification boosted by modified metaheuristic optimization. In: International Conference on Advances in Data-driven Computing and Intelligent Systems. pp. 289–303. Springer (2023)
45. Milutinovic, N., Gajevic, M., Krstovic, J., Petrovic, A., Bacanin, N., Antonijevic, M.: Performance of arithmetic optimization algorithm for elm tuning applied to iot security. In: 2022 30th Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2022)
46. Minic, A., Jovanovic, L., Bacanin, N., Stoean, C., Zivkovic, M., Spalevic, P., Petrovic, A., Dobrojevic, M., Stoean, R.: Applying recurrent neural networks for anomaly detection in electrocardiogram sensor data. Sensors **23**(24), 9878 (2023)
47. Mirjalili, S., Mirjalili, S.: Genetic algorithm. Evolutionary algorithms and neural networks: Theory and applications pp. 43–55 (2019)
48. Mohammed, H., Rashid, T.: Fox: a fox-inspired optimization algorithm. Applied Intelligence **53**(1), 1030–1050 (2023)
49. Petrovic, A., Antonijevic, M., Strumberger, I., Budimirovic, N., Savanovic, N., Janicijevic, S.: Intrusion detection by xgboost model tuned by improved multi-verse optimizer. In: Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022). vol. 104, p. 203. Springer Nature (2023)
50. Petrovic, A., Antonijevic, M., Strumberger, I., Jovanovic, L., Savanovic, N., Janicijevic, S.: The xgboost approach tuned by tlb metaheuristics for fraud detection. In: Proceedings of the 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022). vol. 104, p. 219. Springer Nature (2023)
51. Petrovic, A., Bacanin, N., Zivkovic, M., Marjanovic, M., Antonijevic, M., Strumberger, I.: The adaboost approach tuned by firefly metaheuristics for fraud detection. In: 2022 IEEE World Conference on Applied Intelligence and Computing (AIC). pp. 834–839. IEEE (2022)
52. Petrovic, A., Damaševičius, R., Jovanovic, L., Toskovic, A., Simic, V., Bacanin, N., Zivkovic, M., Spalević, P.: Marine vessel classification and multivariate trajectories forecasting using metaheuristics-optimized extreme gradient boosting and recurrent neural networks. Applied Sciences **13**(16), 9181 (2023)
53. Petrovic, A., Jovanovic, L., Zivkovic, M., Bacanin, N., Budimirovic, N., Marjanovic, M.: Forecasting bitcoin price by tuned long short term memory model. In: 1st International conference on innovation in information technology and business (ICIITB 2022). pp. 187–202. Atlantis Press (2023)
54. Petrovic, A., Strumberger, I., Antonijevic, M., Jovanovic, D., Mladenovic, D., Chabbra, A.: Firefly-xgboost approach for pedestrian detection. In: 2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC). pp. 197–202. IEEE (2022)
55. Petrovic, A., Strumberger, I., Bezdan, T., Jassim, H.S., Nassor, S.S.: Cryptocurrency price prediction by using hybrid machine learning and beetle antennae search approach. In: 2021 29th Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2021)
56. Predić, B., Jovanovic, L., Simic, V., Bacanin, N., Zivkovic, M., Spalevic, P., Budimirovic, N., Dobrojevic, M.: Cloud-load forecasting via decomposition-aided attention recurrent neural network tuned by modified particle swarm optimization. Complex & Intelligent Systems pp. 1–21 (2023)

57. Radomirovic, B., Petrovic, A., Zivkovic, M., Njegus, A., Budimirovic, N., Bacanin, N.: Efficient spam email classification logistic regression model trained by modified social network search algorithm. In: Computational Intelligence and Blockchain in Complex Systems, pp. 39–55. Elsevier (2024)
58. Salb, M., Bacanin, N., Zivkovic, M., Antonijevic, M., Marjanovic, M., Strumberger, I.: Extreme learning machine tuning by original sine cosine algorithm. In: 2022 IEEE World Conference on Applied Intelligence and Computing (AIC). pp. 143–148. IEEE (2022)
59. Salb, M., Jovanovic, L., Bacanin, N., Antonijevic, M., Zivkovic, M., Budimirovic, N., Abualigah, L.: Enhancing internet of things network security using hybrid cnn and xgboost model tuned via modified reptile search algorithm. Applied Sciences **13**(23), 12687 (2023)
60. Savanović, N., Toskovic, A., Petrovic, A., Zivkovic, M., Damaševičius, R., Jovanovic, L., Bacanin, N., Nikolic, B.: Intrusion detection in healthcare 4.0 internet of things systems via metaheuristics optimized machine learning. Sustainability **15**(16), 12563 (2023)
61. Stankovic, M., Gavrilovic, J., Jovanovic, D., Zivkovic, M., Antonijevic, M., Bacanin, N., Stankovic, M.: Tuning multi-layer perceptron by hybridized arithmetic optimization algorithm for healthcare 4.0. Procedia Computer Science **215**, 51–60 (2022)
62. Stoean, C., Zivkovic, M., Bozovic, A., Bacanin, N., Strulak-Wójcikiewicz, R., Antonijevic, M., Stoean, R.: Metaheuristic-based hyperparameter tuning for recurrent deep learning: application to the prediction of solar energy generation. Axioms **12**(3), 266 (2023)
63. Strumberger, I., Bacanin, N., Tuba, M.: Enhanced firefly algorithm for constrained numerical optimization. In: 2017 IEEE congress on evolutionary computation (CEC). pp. 2120–2127. IEEE (2017)
64. Strumberger, I., Beko, M., Tuba, M., Minovic, M., Bacanin, N.: Elephant herding optimization algorithm for wireless sensor network localization problem. In: Technological Innovation for Resilient Systems: 9th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2018, Costa de Caparica, Portugal, May 2-4, 2018, Proceedings 9. pp. 175–184. Springer (2018)
65. Todorovic, M., Petrovic, A., Toskovic, A., Zivkovic, M., Jovanovic, L., Bacanin, N.: Multivariate bitcoin price prediction based on lstm tuned by hybrid reptile search algorithm. In: 2023 16th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS). pp. 195–198. IEEE (2023)
66. Todorovic, M., Stanisic, N., Zivkovic, M., Bacanin, N., Simic, V., Tirkolaee, E.B.: Improving audit opinion prediction accuracy using metaheuristics-tuned xgboost algorithm with interpretable results through shap value analysis. Applied Soft Computing **149**, 110955 (2023)
67. Toskovic, A., Petrovic, A., Jovanovic, L., Bacanin, N., Zivkovic, M., Dobrojevic, M.: Marine vessel trajectory forecasting using long short-term memory neural networks optimized via modified metaheuristic algorithm. In: International Conference on Trends in Sustainable Computing and Machine Intelligence. pp. 51–66. Springer (2023)
68. Tuba, M., Bacanin, N.: Jpeg quantization tables selection by the firefly algorithm. In: 2014 International Conference on Multimedia Computing and Systems (ICMCS). pp. 153–158. IEEE (2014)

69. Vesic, A., Marjanovic, M., Petrovic, A., Strumberger, I., Tuba, E., Bezdan, T.: Optimizing extreme learning machine by animal migration optimization. In: 2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC). pp. 261–266. IEEE (2022)
70. Wang, D., Tan, D., Liu, L.: Particle swarm optimization algorithm: an overview. Soft computing **22**, 387–408 (2018)
71. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE transactions on evolutionary computation **1**(1), 67–82 (1997)
72. Zivkovic, M., Bacanin, N., Antonijevic, M., Nikolic, B., Kvascev, G., Marjanovic, M., Savanovic, N.: Hybrid cnn and xgboost model tuned by modified arithmetic optimization algorithm for covid-19 early diagnostics from x-ray images. Electronics **11**(22), 3798 (2022)
73. Zivkovic, M., Bacanin, N., Zivkovic, T., Strumberger, I., Tuba, E., Tuba, M.: Enhanced grey wolf algorithm for energy efficient wireless sensor networks. In: 2020 zooming innovation in consumer technologies conference (ZINC). pp. 87–92. IEEE (2020)
74. Zivkovic, M., Jovanovic, L., Bacanin, N., Petrovic, A., Savanovic, N., Dobrojevic, M.: Xgboost tuned by hybridized sca metaheuristics for intrusion detection in healthcare 4.0 iot systems. In: International Conference on Engineering, Applied Sciences and System Modeling. pp. 1–16. Springer (2017)
75. Zivkovic, M., Jovanovic, L., Ivanovic, M., Bacanin, N., Strumberger, I., Joseph, P.M.: Xgboost hyperparameters tuning by fitness-dependent optimizer for network intrusion detection. In: Communication and intelligent systems: Proceedings of ICCIS 2021, pp. 947–962. Springer (2022)
76. Zivkovic, M., Jovanovic, L., Ivanovic, M., Krdzic, A., Bacanin, N., Strumberger, I.: Feature selection using modified sine cosine algorithm with covid-19 dataset. In: Evolutionary computing and mobile sustainable networks: Proceedings of ICECMSN 2021, pp. 15–31. Springer (2022)
77. Zivkovic, M., Petrovic, A., Bacanin, N., Djuric, M., Vesic, A., Strumberger, I., Marjanovic, M.: Training logistic regression model by hybridized multi-verse optimizer for spam email classification. In: Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2. pp. 507–520. Springer (2023)
78. Zivkovic, M., Petrovic, A., Bacanin, N., Milosevic, S., Veljic, V., Vesic, A.: The covid-19 images classification by mobilenetv3 and enhanced sine cosine metaheuristics. In: Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2022, pp. 937–950. Springer Nature Singapore Singapore (2022)
79. Zivkovic, M., Petrovic, A., Venkatachalam, K., Strumberger, I., Jassim, H.S., Bacanin, N.: Novel chaotic best firefly algorithm: Covid-19 fake news detection application. In: Advances in Swarm Intelligence: Variations and Adaptations for Optimization Problems, pp. 285–305. Springer International Publishing Cham (2022)
80. Zivkovic, M., Stoean, C., Chhabra, A., Budimirovic, N., Petrovic, A., Bacanin, N.: Novel improved salp swarm algorithm: An application for feature selection. Sensors **22**(5), 1711 (2022)
81. Zivkovic, M., Stoean, C., Petrovic, A., Bacanin, N., Strumberger, I., Zivkovic, T.: A novel method for covid-19 pandemic information fake news detection based on the arithmetic optimization algorithm. In: 2021 23rd international symposium on symbolic and numeric algorithms for scientific computing (SYNASC). pp. 259–266. IEEE (2021)

82. Zivkovic, M., Vesic, A., Bacanin, N., Strumberger, I., Antonijevic, M., Jovanovic, L., Marjanovic, M.: An improved animal migration optimization approach for extreme learning machine tuning. In: International Conference on Intelligent and Fuzzy Systems. pp. 3–13. Springer (2022)
83. Živković, M., et al.: Plant classification using firefly algorithm and support vector machine. In: Sinteza 2022-International Scientific Conference on Information Technology and Data Related Research. pp. 443–448. Singidunum University (2022)
84. Zivkovic, T., Nikolic, B., Simic, V., Pamucar, D., Bacanin, N.: Software defects prediction by metaheuristics tuned extreme gradient boosting and analysis based on shapley additive explanations. Applied Soft Computing **146**, 110659 (2023)