Proceedings of CECNet 2022
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220579

Stability Evaluation of Computational Intelligence-Based Subset Feature Selection Methods on Breast Cancer Data Analysis

Shaode YU [a], Boji LIU [a], Bingjie LI [a], Mingxue JIN [a], Junjie WU [b], and Hang YU [c,1]

[a] *School of Information and Communication Engineering, Communication University of China, Chaoyang, Beijing, China*

[b] *Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, Texas, United States*

[c] *School of Aerospace Science and Technology, Xidian University, Xi'an, Shanxi, China*

Abstract. The stability of computational intelligence based subset feature selection (CI-SFS) has not been explored. In this study, 44 methods are evaluated on BCDR-F03 using 5 stability estimators. Experimental results identify 3 methods achieving 0.55 or higher scores from two estimators, 7 methods leading to good classification (area under the curve $\geq$ 0.80) and 4 potential signatures helping cancer diagnosis. Conclusively, most of the CI-SFS methods seem sensitive to data perturbation and different estimators cause inconsistent results. In future work, attention should be paid to developing robust fitness functions to enhance feature preference and designing advanced estimators to quantify the feature selection stability.

Keywords. Stability, computational intelligence, subset feature selection, breast cancer diagnosis, signature discovery

1. Introduction

Own to the dramatic increase of variable dimension, feature selection (FS) is growingly important in pattern analysis [1–3]. To choose most relevant features, computational intelligence based subset feature selection (CI-SFS) algorithms have been developed [4], and their purpose is to imitate swarming behaviour, social hierarchy, foraging strategy and hunting mechanisim to select a subset of features for user preference.

This study investigates CI-SFS stability on feature preference. Stability is important in machine learning, since it is correlated with experiment-level repeatability and pattern analysis [5]. Meanwhile, CI-SFS has made big progress in the past decades [4]. Thus, it is meaningful to present an evaluation of CI-SFS stability.

Few studies concern FS stability. In [6], algorithms are analyzed using correlation coefficient and Jaccard index. In [7], stability is assessed using two similarity-based es-

[1] Corresponding Author: Hang Yu, Xidian University, Shanxi, China; E-mail: hyu@xidian.edu.cn.

timators. In [8], stability is estimated via adapted Tanimoto distance and correlation coefficients. In [9], stability is quantified via relative weighted consistency and correlation-based measures. In [10], 23 FR algorithms are evaluted using an advanced estimator. These studies pave the way for understanding the FS stability.

In this study, using 5 estimators, we investigate the stability of 44 CI-SFS algorithms on the BCDR-F03, a medical dataset with sufficient instances. The contributions of this study come from several points. Above all, the stability of a large number of bio-inspired CI-SFS algorithms is quantified. Secondly, five estimators are used to show experimental cues on estimator application. Thirdly, on BCDR-F03, several potential signatures are discovered that benefit medical image analysis and cancer diagnosis.

2. Materials and Methods

2.1. Data collection

BCDR-F03 [11] includes 230 benign and 176 malignant breast lesions of 736 mammograms[2]. For representation, $p = 17$ features are computed from intensity (i_mean, i_median, i_std_dev, i_max, i_min, i_kurtosis and i_ skewness), shape (s_area, s_perimeter, s_x_center, s_y_center, s_circularity, s_elongation and s_form) and texture (t_contrast, t_correlation and t_entropy). Since 310 cases are imaged twice [12], to avoid one lesion with multiple records, the first one of each lesion is used and 406 feature records remain.

Table 1 shows the dataset, and 141 records of each group are used in the variable selection procedure. Notably, t-test is conducted, and the features (i_min, i_kurtosis and s_x_center) with no significant difference are removed.

Table 1. Summary of the dataset BCDR-F03 used in this study

	benign (train/test)	malignant (train/test)	p	source
BCDR-F03	230 (141/89)	176 (141/35)	17 (14)	mammogram

2.2. CI-SFS algorithms

Forth-four algorithms are evaluated that use different heuristic optimization strategies[3], including artificial bee colony (ABC) [13], artificial butterfly optimization (ABO) [14], ant colony optimization (ACO) [15], ant colony system (ACS) [16], atom search optimization (ASO) [17], bat algorithm (BA) [18], butterfly optimization algorithm (BOA) [19], cuckoo search (CS) [20], crow search algorithm (CSA) [21], differential evolution (DE) [22], equilibrium optimizer (EO) [23], emperor penguin optimizer (EPO) [24], firefly algorithm (FA) [25], fruit fly optimization algorithm (FFOA) [26], flower pollination algorithm (FPA) [27], genetic algorithm (GA) [28], genetic algorithm tournament (GAT) [29], generalized normal distribution optimization (GNDO) [30], gravitational search algorithm (GSA) [31], grey wolf optimizer (GWO) [32], henry gas solubility optimization (HGSO) [33], Harris hawks optimization (HHO) [34], human learning optimization

[2]http://bcdr.inegi.up.pt
[3]https://github.com/JingweiToo/Wrapper-Feature-Selection-Toolbox

(HLO) [35], harmony search (HS) [36], Jaya algorithm (JAYA) [37], Monarch butterfly optimization (MBO) [38], moth-flame optimization (MFO) [39], marine Predators Algorithm (MPA) [40], Manta ray foraging optimization (MRFO) [41], multi-verse optimizer (MVO) [42], poor and rich optimization algorithm (PARO) [43], pathfinder algorithm (PFA) [44], particle swarm optimization (PSO) [45], simulated annealing (SA) [46], satin bowerbird optimizer (SBBO) [47], sine cosine algorithm (SCA) [48], slime mould algorithm (SMA) [49], symbiotic organisms search (SOS) [50], salp swarm algorithm (SSA) [51], tree growth algorithm (TGA) [52], tree-seed algorithm (TSA) [53], whale optimization algorithm (WOA) [54], and weighted superposition attraction (WSA) [55].

2.3. Experiment design

Figure 1 shows stability estimation and classification performance. In each iteration, a dataset $\{(X,y)\}$ is divided for training $\{(X^{train}, y^{train})\}$ and testing $\{(X^{test}, y^{test})\}$, a CI-SFS yields a binary vector $\vec{f_i}$ ($\vec{f_i} =< f_{i,1}, ..., f_{i,k}, ..., f_{i,p} >$) after the i^{th} run of p features. Specifically, $f_{i,k} = 1$ indicates the k^{th} feature is selected, and classification metrics are computed. After $N = 500$ iterations, S values are estimated, and metrics are averaged.

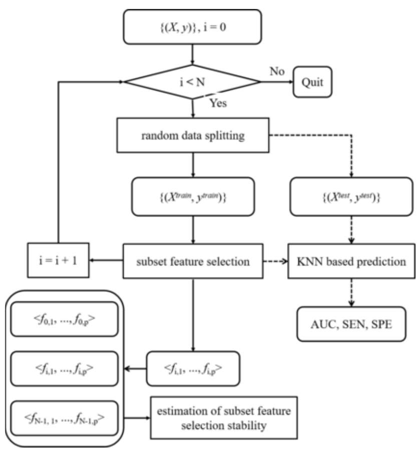

Figure 1. The procedure of estimating stability and classification performance.

2.4. Stability estimator

Five estimators are employed [56]. The computatation of similarity based estimators Jaccard, Dice and Ochi are defined as $J = \frac{\vec{f_i} \cap \vec{f_j}}{\vec{f_i} \cup \vec{f_j}}$, $D = \frac{2 \times (\vec{f_i} \cap \vec{f_j})}{\vec{f_i} + \vec{f_j}}$ and $O = \frac{\vec{f_i} \cap \vec{f_j}}{\sqrt{\vec{f_i} \times \vec{f_j}}}$ respectively, in which $\cap$ and $\cup$ correspond to intersection and union parts of $\vec{f_i}$ and $\vec{f_j}$. To the entropy estimator, it computes the normalized frequency of features, finds out the features with non-zero frequency $\widehat{p}(s_i)$ and quantifies the entropy as $E = -\sum \widehat{p}(s_i) log_2 \widehat{p}(s_i)$. The stability estimator (Nogueria) recasts the stability measure procedure as an estimation of a random variable [5], and allows for reliable comparison across different procedures.

2.5. Classification performance metrics

Three metrics of the area under the receiver operating characteristic curve (AUC), sensitivity (SEN) and specificity (SPE) are used to evaluate the performance of tumor classification [3]. Given the ground truth and predicted labels, AUC reveals the capacity of tumor differentiation based on the curve of prediction probability, SEN reflects the ability of a model to correctly recognize malignant lesions, and SPE shows the ability of a model to identify benign cases correctly. To each metric, a higher value indicates a better performance. In this binary problem ($y \in \{0,1\}$), the label of malignant cases is $y = 1$.

3. Results

3.1. Estimated stability

Figure 2 shows the stability values of CI-SFS algorithms. The horizontal axis lists CI-SFS names, and the vertical axis is S values. It shows 3 algorithms achieve stable feature preference (DE, $S \geq 0.58$; FFOA $S \geq 0.55$; WSA, $S \geq 0.60$) identified by DICE (pink) and Ochi (black). The values using entropy- and Nogueria-based estimators are low.

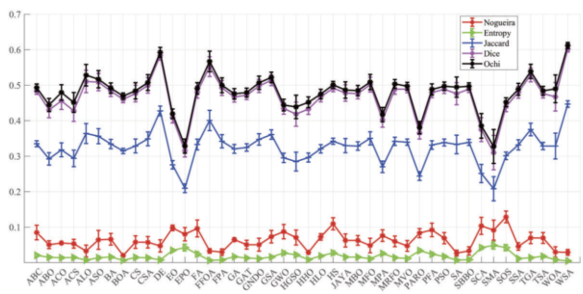

Figure 2. Stability values of CI-SFS algorithms from five estimators.

3.2. Prediction performance

Figure 3 shows the performance of tumor classification. The horizontal axis shows the CI-SFS names, the vertical axis shows the AUC values, and KNN is the classifier. It is observed that CI-SFS algorithms lead to good performance (AUC $\geq$ 0.70) and 7 algorithms (ABC, FA, HLO, HS, MPA, PFA and SOS) achieve AUC $\geq$ 0.80.

3.3. Potential signatures

Figure 4 shows the signagures discovered by CI-SFS algorithms. In the $N = 500$ iterations, when a feature is selected more than 250 time (*i.e.*, $\geq$ 50% chance of selection), it is defined as a potential signature. Further, the number of CI-SFS algorithms that define features as signatures is summarized. It is observed that there are 44, 42, 28 and 34 algorithms that respectively identify s_circularity, s_y_center, t_contr and s_form as the potential signatures in BCDR-F03 data analysis, followed by i_skewness with 15 times of selection, and the other features are selected less than 6 times.

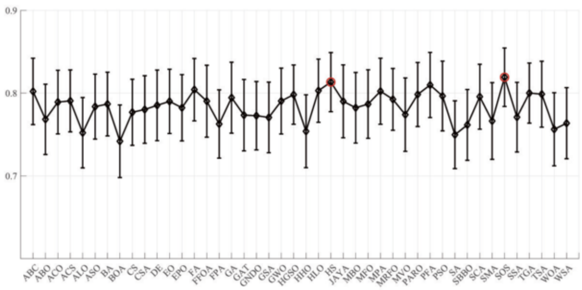

Figure 3. AUC values of CI-SFS-guided classification results.

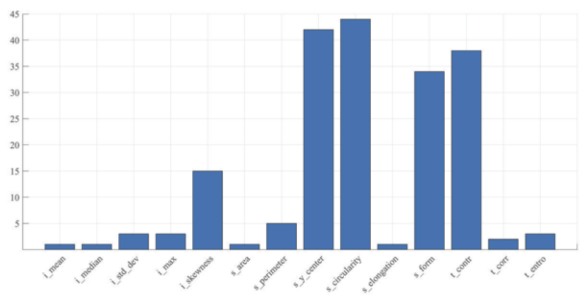

Figure 4. Potential signagures discovered by different CI-SFS algorithms.

4. Discussion

On the BCDR-F03 dataset with 406 samples of lesion cases, up to 44 CI-SFS algorithms are investigated using 5 stability estimators, and the breast tumor classification performance is also explored. All the algorithms achive good prediction with AUC ≥ 0.70, 4 potiential signatures are identified consistently, while only 3 algorithms achieve stable feature preference (S $\geq$ 0.55) when using Dice and Ochi as the estimator.

This study concerns CI-SFS stability. Five previous studies [6–10] explore the stability of FS methods by using different estimators. Notably, [10] focuses on FR stability, 23 methods are explored, while unfortunately, 3 methods generate stable feature ranks when using Nogueria [5] as the estimator. In this study, 44 algorithms have been evaluated, and few methods (DE [22], FFOA [26], WSA [55]) shown in Figure 2 yield robust feature preference when using estimator Dice or Ochi [56]. The studies reveal that stability is an important characteristic and more attention should be paid to this topic.

Two similarity-based estimators (Dice and Ochi) identify three CI-SFS algorithms achieving good stability ($S \geq 0.55$, Figure 2). Firstly, 44 methods are evaluated using 5 estimators. Both the number of methods and estimators surpass that of previous studies [6–10]. Secondly, two similarity-based estimators find 3 stable algorithms. In details, among the five estimators, two estimators reveal the stability values of two methods are less than 0.2, and the Jaccard values of methods show similar value pattern as Dice and Ochi but much lower. It indicates that gaps exist among different estimators that should be well addressed in the future work.

CI-SFS algorithms lead to good prediction (Figure 3). These algorithms result in AUC values larger than 0.70, close to the baseline work [11]. Moreover, ABC [13], FA [25], HLO [35], HS [36], MPA [40], PFA [44] and SOS [50] achieve AUC values larger than 0.80, better than the baseline [11]. It reveals that stability and effectiveness are important yet different characteristics of feature preference.

Moreover, four features are recognized as potential signatures by most CI-SFS algorithms (Figure 4) that may help cancer diagnosis and precision medicine. Among the four features, three features describe shape information (s_circularity, s_y_center and s_form), and one feature quantifies mass lesion texture (t_contr). In clinical practice, the breast imaging-reporting and data system descriptor (BI-RADS) recommends malignant lesions in MAM images are prone to show irregular shapes and inhomogeneous contrast, indicating that signatures discovered in the present study is in accordance to clinical guidelines [57].

Several limitations exist in the present study. Firstly, CI-SFS algorithms are investigated on one dataset, and for comprehensive stability analysis, more medical datasets should be used. Secondly, five estimators are used to quantify the CI-SFS stability, while the results from two estimators are much lower, that may induce controversy among different estimators [5]. Last but not the least, besides handcrafted features, deeply learned features will be studied in our future work to improve network explainability, robustness and generalization capacity [58].

5. Conclusions

Forty-four CI-SFS algorithms have been investigated on the BCDR-F03 dataset by using five stability estimators, and three algorithms are identified consistently exhibiting good stability from two similarity-based estimators, while the gap among different estimators should be considered in estimator design.

References

[1] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, 2018.

[2] Z. Zhang, X. Liang, W. Qin, S. Yu, and Y. Xie, "matFR: a MATLAB toolbox for feature ranking," Bioinformatics, vol. 36(19), pp. 4968–4969, 2020.

[3] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, "A technical review of convolutional neural network-based mammographic breast cancer diagnosis," Computational and mathematical methods in medicine, 2019.

[4] K. Abu, I. Aljarah, A. Sharieh, E. Abd, R. Damaševičius, and T. Krilavičius, "A review of the modification strategies of the nature inspired algorithms for feature selection problem," Mathematics, vol. 10(3), pp. 464, 2022.

[5] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," Journal of Machine Learning Research, vol. 18(1), pp. 6345–6398, 2017.

[6] N. López, M. García-Ordás, F. Vitelli-Storelli, P. Fernández-Navarro, C. Palazuelos, and R. Alaiz-Rodríguez, "Evaluation of feature selection techniques for breast cancer risk prediction," International Journal of Environmental Research and Public Health. vol. 18(20), pp. 10670, 2021.

[7] N. Cueto-López, M. García-Ordás, V. Dávila-Batista, V. Moreno, N. Aragonés, and R. Alaiz-Rodríguez, "A comparative study on feature selection for a risk prediction model for colorectal cancer," Computer methods and programs in biomedicine. vol. 177, pp. 219–229, 2019.

[8] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," Knowledge and information systems. vol. 12(1), pp. 95–116, 2007.

[9] D. Dernoncourt, B. Hanczar, and J. Zucker. "Analysis of feature selection stability on high dimension and small sample data," Computational statistics & data analysis. vol. 71, pp. 681–693, 2014.

[10] S. Yu, B. Li, B. Liu, M. Jin, J. Wu, and H. Yu. "A stablity evaluation of feature ranking algorithms on breast cancer data analysis," Frontiers in Artificial Intelligence and Applications. 2022.

[11] J. Arevalo, F. González, R. Ramos-Pollán, J. Oliveira, and M. Lopez. "Representation learning for mammography mass lesion classification with convolutional neural networks," Computer methods and programs in biomedicine. vol. 127, pp. 248–257, 2016.

[12] S. Yu, L. Liu, Z. Wang, G. Dai, and Y. Xie. "Transferring deep neural networks for the differentiation of mammographic breast lesions," Science China Technological Sciences. vol. 62(3), pp. 441–447, 2019.

[13] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony (ABC) algorithm and applications," Artificial Intelligence Review, vol. 42(1), pp. 21–57, 2014.

[14] D. Rodrigues, V. de Albuquerque, and J. Papa, "A multi-objective artificial butterfly optimization approach for feature selection," Applied Soft Computing, vol. 94, pp. 106442, 2020.

[15] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," IEEE computational intelligence magazine, vol. 1(4), pp. 28–39, 2006.

[16] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem," IEEE Transactions on evolutionary computation, vol. 1(1), pp. 53–66, 1997.

[17] W. Zhao, L. Wang, and Z. Zhang, "Atom search optimization and its application to solve a hydrogeologic parameter estimation problem," Knowledge-Based Systems, vol. 163, pp. 283–304, 2019.

[18] X. Yang and X. He, "Bat algorithm: literature review and applications," International Journal of Bio-inspired computation, vol. 5(3), pp. 141–149, 2013.

[19] S. Arora and S. Singh, "Butterfly optimization algorithm: a novel approach for global optimization," Soft Computing, vol. 23(3), pp. 715–734, 2019.

[20] X. Yang and S. Deb, "Cuckoo search: recent advances and applications," Neural Computing and applications, vol. 24(1), pp. 169–174, 2014.

[21] A. Hussien, M. Amin, M. Wang, G. Liang, A. Alsanad, A. Gumaei, and H. Chen, "Crow search algorithm: theory, recent advances, and applications," IEEE Access, vol. 8, pp. 173548–173565, 2020.

[22] S. Das and P. M. Suganthan, "Differential evolution: A survey of the state-of-the-art," IEEE transactions on evolutionary computation, vol. 15(1), pp. 4–31, 2010.

[23] A. Faramarzi, M. Heidarinejad, B. Stephens, and S. Mirjalili, "Equilibrium optimizer: A novel optimization algorithm," Knowledge-Based Systems, vol. 191, pp. 105190, 2020.

[24] G. Dhiman, and V. Kumar, "Emperor penguin optimizer: a bio-inspired algorithm for engineering problems," Knowledge-Based Systems, vol. 159, pp. 20–50, 2018.

[25] X. Yang, and A. Slowik, "Firefly algorithm," Swarm Intelligence Algorithms, vol. 62, pp. 163–174, 2020.

[26] Q. Pan, H. Sang, J. Duan, and L. Gao, "An improved fruit fly optimization algorithm for continuous function optimization problems," Knowledge-Based Systems, vol. 62, pp. 69–83, 2014.

[27] X. S. Yang, "Flower pollination algorithm for global optimization," International conference on unconventional computing and natural computation, pp. 240–249, 2012.

[28] S. Mirjalili, "Genetic algorithm," Evolutionary algorithms and neural networks, pp. 43–55, 2019.

[29] C. Ding, Y. Cheng, and M. He, "Two-level genetic algorithm for clustered traveling salesman problem with application in large-scale TSPs," Tsinghua Science & Technology, vol. 12(4), pp. 459–465, 2007.

[30] Y. Zhang, Z. Jin, and S. Mirjalili, "Generalized normal distribution optimization and its applications in parameter extraction of photovoltaic models," Energy Conversion and Management, vol. 224, pp. 113301, 2020.

[31] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "GSA: a gravitational search algorithm," Information sciences, vol. 179(13), pp. 2232–2248, 2009.

[32] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in engineering software, vol. 69, pp. 46–61, 2014.

[33] F. A. Hashim, E. H. Houssein, M. S. Mabrouk, W. Al-Atabany, and S. Mirjalili, "Henry gas solubility optimization: A novel physics-based algorithm," Future generation computer systems, vol. 101, pp. 646–667, 2019.

[34] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," Future generation computer systems, vol. 97, pp. 849–872, 2019.

[35] L. Wang, R. Yang, H. Ni, W. Ye, M. Fei, and P. M. Pardalos, "A human learning optimization algorithm and its application to multi-dimensional knapsack problems," Applied Soft Computing, vol. 34, pp.

736–743, 2015.
[36] Z. W. Geem, J. H. Kim, and G. V. Loganathan, "A new heuristic optimization algorithm: harmony search," Simulation, vol. 76(2), pp. 60–68, 2001.
[37] E. H. Houssein, A. G. Gad, and Y. M. Wazery, "Jaya algorithm and applications: A comprehensive review," Metaheuristics and Optimization in Computer and Electrical Engineering, pp. 3–24, 2021.
[38] G. Wang, S. Deb, and Z. Cui, "Monarch butterfly optimization," Neural computing and applications, vol. 31(7), pp. 1995–2014, 2019.
[39] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," Knowledge-based systems, vol. 89, pp. 228–249, 2015.
[40] A. Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine Predators Algorithm: A nature-inspired metaheuristic," Expert systems with applications, vol. 152, pp. 113377, 2020.
[41] W. Zhao, Z. Zhang, and L. Wang, "Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications," Engineering Applications of Artificial Intelligence, vol. 87, pp. 103300, 2020.
[42] S. Mirjalili, S. H. Mirjalili, and A. Hatamlou, "Multi-verse optimizer: a nature-inspired algorithm for global optimization," Neural Computing and Applications, vol. 27(2), pp. 495–513, 2016.
[43] S. H. S. Moosavi and V. K. Bardsiri, "Poor and rich optimization algorithm: A new human-based and multi populations algorithm," Engineering Applications of Artificial Intelligence, vol. 86, pp. 165–181, 2019.
[44] H. Yapici and N. Cetinkaya, "A new meta-heuristic optimizer: Pathfinder algorithm," Applied soft computing, vol. 78, pp. 545–568, 2019.
[45] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," Swarm intelligence, vol. 1(1), pp. 33–57, 2007.
[46] D. Bertsimas and J. Tsitsiklis, "Simulated annealing," Statistical science, vol. 8(1), pp. 10–15, 1993.
[47] S. H. S. Moosavi and V. K. Bardsiri, "Satin bowerbird optimizer: A new optimization algorithm to optimize ANFIS for software development effort estimation," Engineering Applications of Artificial Intelligence, vol. 60, pp. 1–15, 2017.
[48] S. Mirjalili, "SCA: a sine cosine algorithm for solving optimization problems," Knowledge-based systems, vol. 96, pp. 120–133, 2016.
[49] S. Li, H. Chen, M. Wang, A. Heidari, and S. Mirjalili, "Slime mould algorithm: A new method for stochastic optimization," Future Generation Computer Systems, vol. 111, pp. 300–323, 2020.
[50] M. Cheng and D. Prayogo, "Symbiotic organisms search: a new metaheuristic optimization algorithm," Computers & Structures, vol. 139, pp. 98–112, 2014.
[51] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," Advances in engineering software, vol. 114, pp. 163–191, 2017.
[52] A. Cheraghalipour, M. Hajiaghaei-Keshteli, and M. M. Paydar, "Tree Growth Algorithm (TGA): A novel approach for solving optimization problems," Engineering Applications of Artificial Intelligence, vol. 72, pp. 393–414, 2018.
[53] M. S. Kiran, "TSA: Tree-seed algorithm for continuous optimization," Expert Systems with Applications, vol. 42(19), pp. 6686–6698, 2015.
[54] S. Mirjalili and A. Lewis, "The whale optimization algorithm," Advances in engineering software, vol. 95, pp. 51–67, 2016.
[55] A. Baykasoğlu and S. Akpinar, "Weighted Superposition Attraction (WSA): A swarm intelligence algorithm for optimization problems–Part 2: Constrained optimization," Applied Soft Computing, vol. 37, pp. 396–415, 2015.
[56] A. Bommert and M. Lang, "stabm: Stability measures for feature selection," Journal of Open Source Software, vol. 6(59), pp. 3010, 2021.
[57] D. Spak, J. Plaxco, L. Santiago, M. Dryden, and B. Dogan, "Bi-rads fifth edition: A summary of changes," Diagnostic and interventional imaging, vol. 98, pp. 179–190, 2017.
[58] S. Yu, M. Chen, E. Zhang, J. Wu, H. Yu, Z. Yang, L. Ma, X. Gu, and W. Lu, "Robustness study of noisy annotation in deep learning based medical image segmentation," Physics in Medicine & Biology, vol. 65(17), pp. 175007, 2020.