University of Thi-Qar Journal for Engineering Sciences
Volume 14, Issue 1, August 2024, Published Online: 06 August 2024
https://doi.org/10.31663/utjes.14.1.541

A Hybrid Feature Selection Approach Based on Firefly Algorithm and Simulated Annealing for Cancer Datasets

Hadeel Tariq Ibrahim[1] • Wamidh Jalil Mazher [2] • Zainab Fadhil Yaseen [3]

Received: 9th November 2023 / Accepted: 31st January 2024

Abstract

The recent tendency of research is to hybridize two or more metaheuristics algorithms to find superior solutions in the field of feature selection problems. The Firefly Algorithm (FA) is a population-based optimization algorithm that attempts to replicate the normal behavior of firefly insects in seeking food. FA is broadly used in numerous engineering fields, but it endures from certain restrictions. This study emphasis emphasizes hybridizing FA with the Simulated Annealing algorithm (SA) as a strong local search algorithm to overcome FA limits and improve the overall performance in feature selection. In other words, a high-level relay hybrid (HRH) model is proposed in which self-contained optimization (i.e., FA and SA) are implemented in sequence. Obviously, metaheuristic algorithms (like FA) are not suitable for fine adjustment structures that are so near to optimal solutions, whereas local search algorithms (like SA) are the opposite. Accordingly, in the proposed FASA+FS model, the best regions are located by FA and then inputted to SA, respectively.

Keywords— Optimization, Metaheuristics, Firefly Algorithm, Simulated Annealing algorithm, Hybrid approach, Feature selection, Cancer datasets.

1 Introduction

Metaheuristics have been studied in a variety of contexts, including feature selection. Because a comprehensive search leans towards producing all viable solutions for a problem, metaheuristics outperform exact search techniques (Guyon & De, 2003). Even though the metaheuristic strategy does not ensure that the optimal solution will be found in every run, it does guarantee that a suitable solution will be found in a fair amount of time (E.-G. Talbi, 2009). Intensification and diversification, sometimes referred to as exploitation and exploration, are the two main constituents of any meta-heuristic algorithm among their many other parts. Metaheuristic algorithms must produce a wide range of solutions employing diversification or exploration approaches in order to explore the search space globally (Zhang et al., 2016). In the literature, many metaheuristics such as Simulated Annealing (SA) (Busetti, 2003), Particle Swarm Optimization (PSO) (Kennedy & Eberhart, n.d.), Differential Evolution (DE) (Khushaba et al., 2008), and Ant Colony Optimization (ACO) (Kabir et al., 2013), Grasshopper Optimization Algorithm (GOA) (Ibrahim et al., 2019), Salp Optimization Algorithm(Tariq

Hadeel Tariq Ibrahim
Hadeel.tariq@shu.edu.iq

Wamidh Jalil Mazher
wamidh.mazher@stu.edu.iq

Zainab Fadhil Yaseen
zyaseen@utq.edu.iq

[1] Information Technology Division, Al-Shatrah University, Al-Shatrah, Iraq

[2] Electrical Engineering Dept., Southern Technical University, Basra, Iraq

[3] University of Thi-Qar, Nassirya, Iraq

Ibrahim et al., 2017), Harriss Hawks Optimizer (HHO) (Ibrahim et al., 2023) have been utilized in searching feature subsets space for choosing (sub) optimal set of features. Our goal in this study is to use the FireFly algorithm (FA) and the SA to create a fresh fusion wrapper technique to improve the efficiency of the feature selection task.

According to (E. G. Talbi, 2002), metaheuristic hybridization paradigms are categorized into high and low levels based on the sequence of employing their functions. At a high level, there is no direct connection between the inner workings of combined algorithms (self-contained), whereas, at a low level, each metaheuristic function can be substituted by the other algorithm function. Relay and teamwork hybridization are two mechanisms that can be used at both the high and low levels. During relay hybridization, a collection of metaheuristics is applied in a workflow style, in which the second algorithm uses (inputs) the output of the first, whereas, in teamwork style, many collaborating agents boost at the same time; all agents search for a solution space (E. G. Talbi, 2002). In this paper, we aim to use the high-level relay hybrid (HRH) layout for feature selection issues. The FA algorithm (population-based) will be hybridized with the SA algorithm (single solution-based), and in this case, the SA algorithm will enhance the exploitation of the FA algorithm. The Firefly algorithm is presented by (Yang, 2010). The flashing conduct of fireflies stimulates the algorithm. The first rule employed to build the algorithm is that an individual firefly will be magnetized to an illuminator firefly, and if there is no illuminator firefly, it will move aimlessly. The main disadvantage of the FA algorithm is that it requires a suitable setting for parameters and a high number of iterations (Zhang et al., 2016); consequently, it needs a long runtime in high-dimensional datasets (Mohammadiyan & Ghadi, 2017). The strengths of the FA and SA algorithms can be merged to create a hybrid model that outperforms each alone. This hybridization is intended to increase both the exploitation time and the efficacy of the FA algorithm. The suggested technique seeks to boost the exploitation of the FA algorithm through the use of the SA algorithm in the high-level relay hybrid (HRH) style.The remainder of this research is structured as follows: section 2 lists the closest related works to this study, and section 3 delves into the specifics of the proposed approach. Section 4 contains the experimental results. Finally, conclusions and future work are considered in Section 5.

2 The related works

In the feature selection area, numerous hybrid metaheuristic paradigms have been suggested successfully. In (Oh et al., 2004), a hybrid paradigm was first proposed by hybridizing genetic algorithms for feature selection in which GA weakness in fine-tuning close to local optimum agents had been overcome by hybridization style. Artificial bee colony (ABC) optimization and differential evolution (DE) algorithms were hybridized for feature selection (ZorarpacI and Özel, 2016). In such a study, the proposed hybrid technique was assessed by employing 15 datasets from the UCI Repository. Mafarja and Mirjalili proposed two hybridization paradigms to solve feature selection problems by combining the Whale Optimization Algorithm (WOA) and Simulated Annealing (SA) algorithm (Mafarja & Mirjalili, 2017).

3 The proposed algorithm

As a binary optimization problem, feature selection has only two possible solutions: binary 0 or 1. A binary version of the FA algorithm is created before employing it with the feature selection issue. The number of features in the initial data set determines the overall length of the vector, which is used in this study to represent the outcome as a one-dimensional vector. Each vector (cell) value is represented by one or zero. The number 1 indicates that the proper attribute was selected; if not, the value is set to 0, and so on. Each solution is evaluated using the recommended fitness function based on the KNN classifier to establish its classification accuracy and the number of selected features. In this technique, the instances closest to the assess instance are determined, and an easy model is built on the list of k nearest neighbors to find the class label (Kaya Keles et al., 2021).

The fitness function described in Eq. (1) is used to evaluate search agents in both FA and SA algorithms as a way to equilibrium the total number of selected attributes in every possible solution (or minimum) and classification-accuracy (or maximum).

$$Fit = AB_R(D) + C\frac{|R|}{|No|} \qquad (1)$$

where $B_R(D)$ denotes the error rate for classification of a given classifier (the KNN classifier is employed here). In addition, $|R|$ is the cardinal of the chosen subset, and $|No|$ is the entire number of attributes in the dataset, A and B are a pair of parameters that represent the status of classification level and subset length, A ϵ[0, 1] and C = (1 -A), respectively, implemented from (Zawbaa et al., 2016).

The FASA+FS algorithm is proposed in this study based on High-level relay hybridization (HRH) model (E. G. Talbi, 2002). According to HRH model, developed binary FA is applied to locate the best solution, then the output of FA is inputted to SA to improve the optimum selected feature. Fig.1 visualized the flow diagram for FASA+FS algorithm. The main steps of such an algorithm are categorized into:

A. **Preprocessing:**
   1- Normalizing Data: The prior processing activity in feature selection is available. In the [0,1] phase, features are normalized to be restricted.
   2- Creating training and testing sets: We divided each of the cancer datasets into two groups: training and testing. The training set constituted 80% of the total dataset in the binary FA algorithm, with the remaining 20% being used for the testing set. We used the KNN classifier to run the training and testing sets in order to create the model (Sun et al., 2023).
   3- Selecting a subset of features: In this case, the training set's features with values of 1 have been selected.

B. **Phase#1:**
   1- Fitness assessment: The vectors from the designated training set were utilized to train the KNN classifier, and as a result, the classification performance was determined using Equation (1).
   2- Stopping condition: By identifying the top iteration, the entire process has been halted. In actuality, the maximum repetition was fixed at 5.
   3- Performing binary FA algorithm

C. **Phase#2**
   In this phase, the resulting subset of features from Phase 1 is inputted into the SA algorithm. Combining SA with FA through FASA+FS improves FA's capacity to explore and search more accurately during the last evolutionary stage and allows FA to break out of the local optimum (Pan et al., 2019). At the end of this phase, we obtain the final optimum subset of features.

4 Experiments

Datasets: MATLAB is used to implement the proposed algorithm. For the years 2010, 2011, and 2012, we used government biomedical datasets for Iraqi cancer patients. Such datasets are not public and require official authentication. The data sets needed to be cleaned up, and their bias and irregular values, which impact classification performance, were removed because they contained some noise. Following pre-processing, we were able to acquire clean datasets with hundreds of instances and 16 attributes.

To assess the effectiveness of the suggested strategies, experiments are conducted using 14 FS benchmark datasets for actual cancer datasets for 14 different forms of cancer in Iraq (2010–2012) (Ministry of Health, 2017). The datasets used are detailed in Table 1, including the number of features and instances in every dataset.

Table 1: List of experimental datasets (Ministry of Health, 2017)

No.	Data set	Total number of instances	Total Number of features
1	Abdomen cancer	471	16
2	Bladder cancer	4288	16
3	Blood cancer	4788	16
4	Bones cancer	950	16
5	Brain cancer	2935	16
6	Colon cancer	3258	16
7	Eye cancer	179	16
8	Glands cancer	1655	16

9	Heart cancer	183	16
10	Liver cancer	2842	16
11	Lungs cancer	4984	16
12	Lymph cancer	5448	16
13	Naso cancer	1818	16
14	Nerve cancer	1175	16

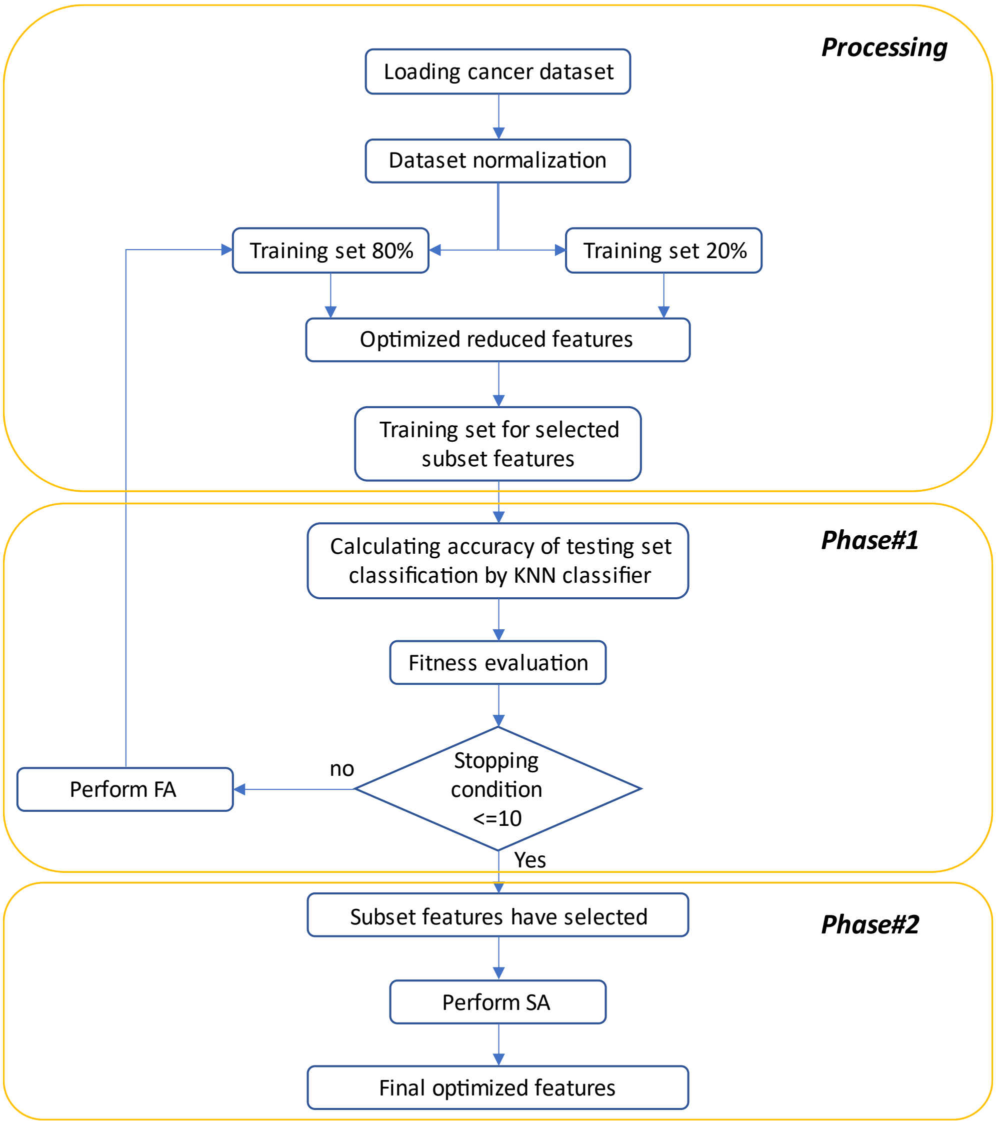


Figure 1: The flow diagram for proposed FASA+FS algorithm

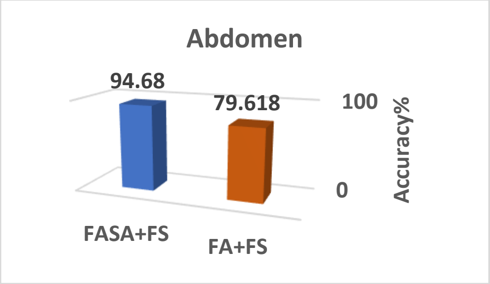
5 Results & Discussions

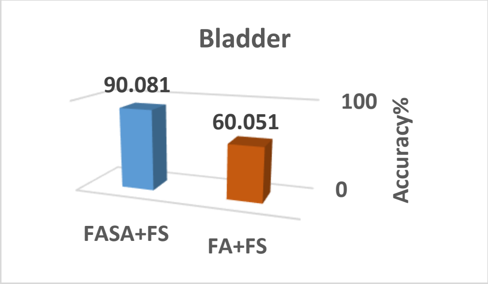
Here, the efficiency of the proposed FASA+FS is evaluated based on two standards: the accuracy of feature selection and the number of selected (resulted) features. Furthermore, FASA+FS results based on the two mentioned metrics are compared with the FA+FS algorithm applied to the same datasets.

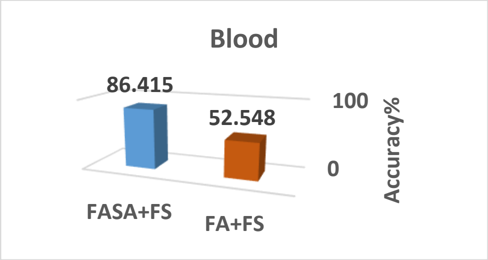
Accuracy of feature selection: Table 2 compares the feature-selection accuracy of the FA+FS and FASA+FS algorithms. In such a table, accuracies are visualized in Fig.2. It is obvious that FASA+FS outdone FA+FS algorithms applied on 14 datasets with average accuracy (93.511%) versus (68.635%) by the FA+FS algorithm.

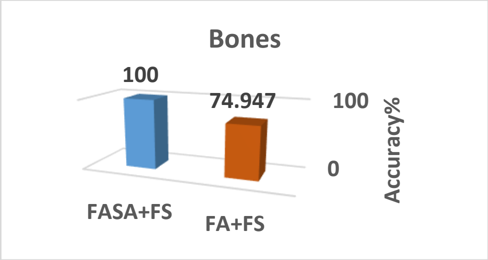
Table 2: Comparison between FA+FS and FASA+FS accuracies (Ministry of Health, 2017)

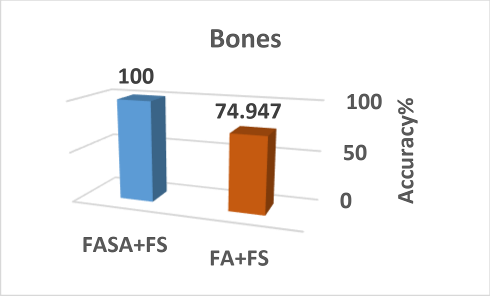
No.	Data set	FA + FS	FASA + FS
1	Abdomen cancer	79.617	94.680
2	Bladder cancer	60.051	90.081
3	Blood cancer	52.548	86.415
4	Bones cancer	74.947	100
5	Brain cancer	68.586	87.734
6	Colon cancer	52.302	87.096
7	Eye cancer	61.452	97.142
8	Glands cancer	79.516	93.353
9	Heart cancer	63.934	97.222
10	Liver cancer	52.076	94.014
11	Lungs cancer	88.643	99.899
12	Lymph cancer	75.642	93.939
13	Naso cancer	78.382	93.112
14	Nerve cancer	73.191	94.468
Average accuracy		68.635	93.511

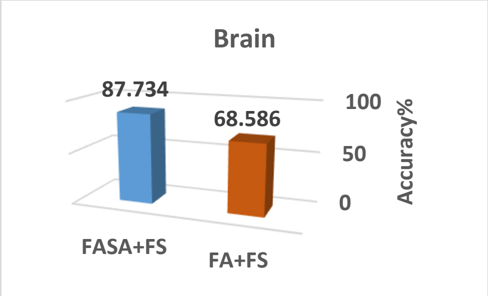


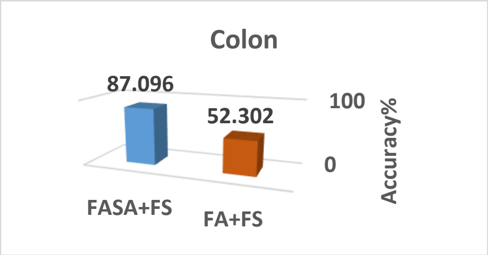


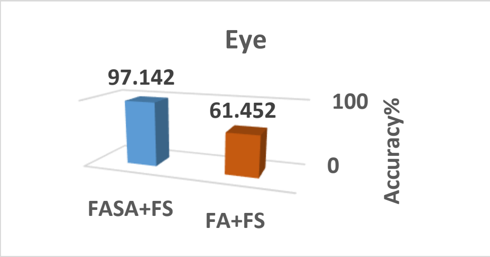












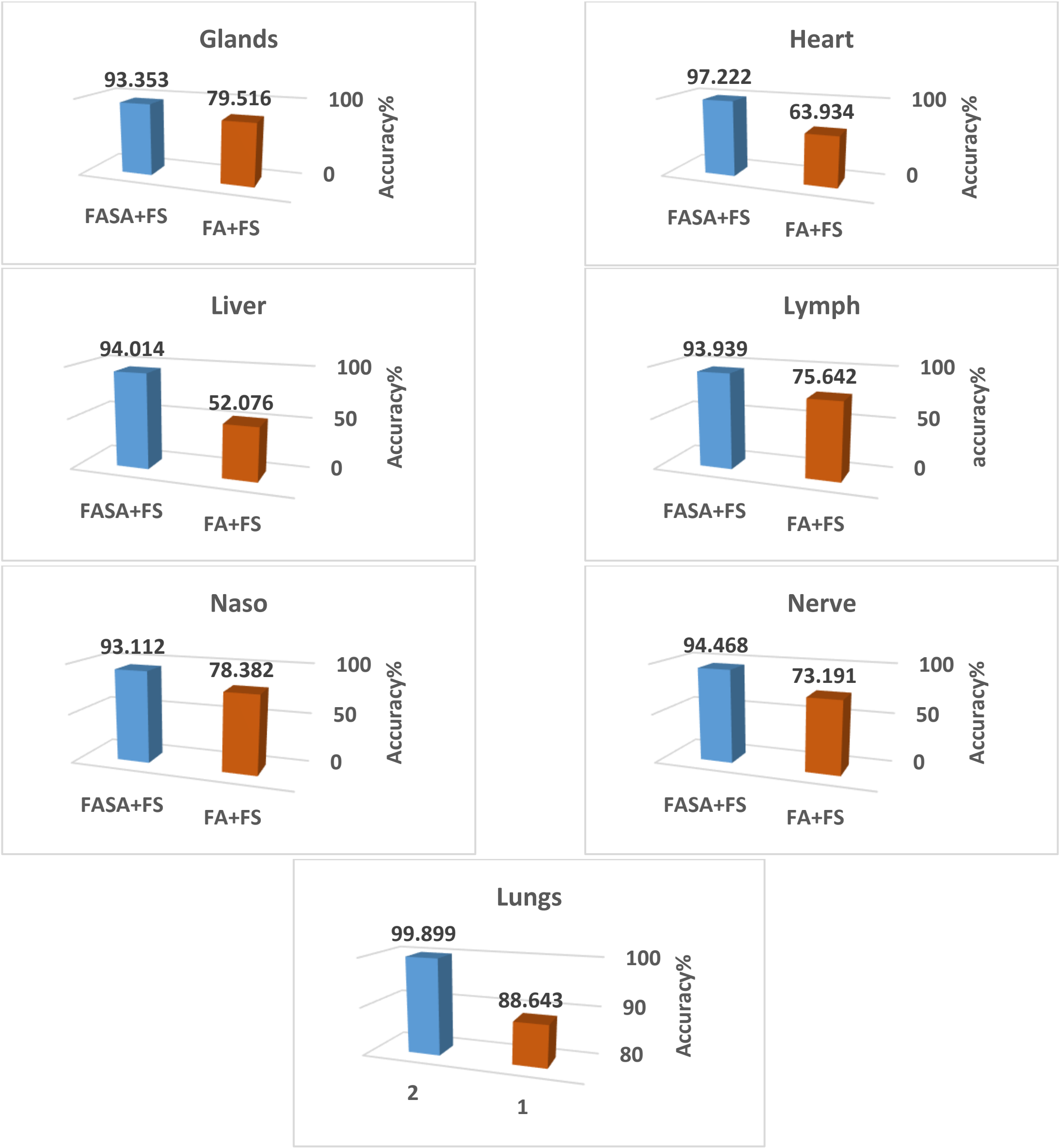


Figure 2: FASA+FS versus FA+FS accuracies

Count of selected features: The first-class classifying algorithm must be able to outdo the one with the lowest classification error rate by choosing the fewest features (Aljarah et al., 2020). The FASA+FS algorithm determines the lowest number of selected features in Table 3 (shown in Fig.3). On 14 datasets, FASA+FS exceeded the other algorithm. The average no. of selected features by FA+FS and FASA+FS algorithms is depicted in Fig.4, and it depicted that a lower average no. of selected features is obtained by FASA+FS (3.357) versus (7.143) by FA+FS. By permitting hill-climbing movements in the hopes of discovering a global optimum, SA is used to avoid local optima (Pan et al., 2019), which is the main reason FASA+FS achieves the lowest average no. of selected features.

Table 3: Comparison of the number of chosen attributes in FA+FS and FASA+FS algorithms

No.	Data set	FA + FS	FASA + FS
1	Abdomen	9	3
2	Bladder	8	3
3	Blood	8	5
4	Bones	6	5
5	Brain	8	2
6	Colon	7	5

7	Eye	6	1
8	Glands	8	4
9	Heart	8	1
10	Liver	5	4
11	Lungs	6	1
12	Lymph	8	6
13	Naso	6	5
14	Nerve	7	2
Average		7.143	3.357

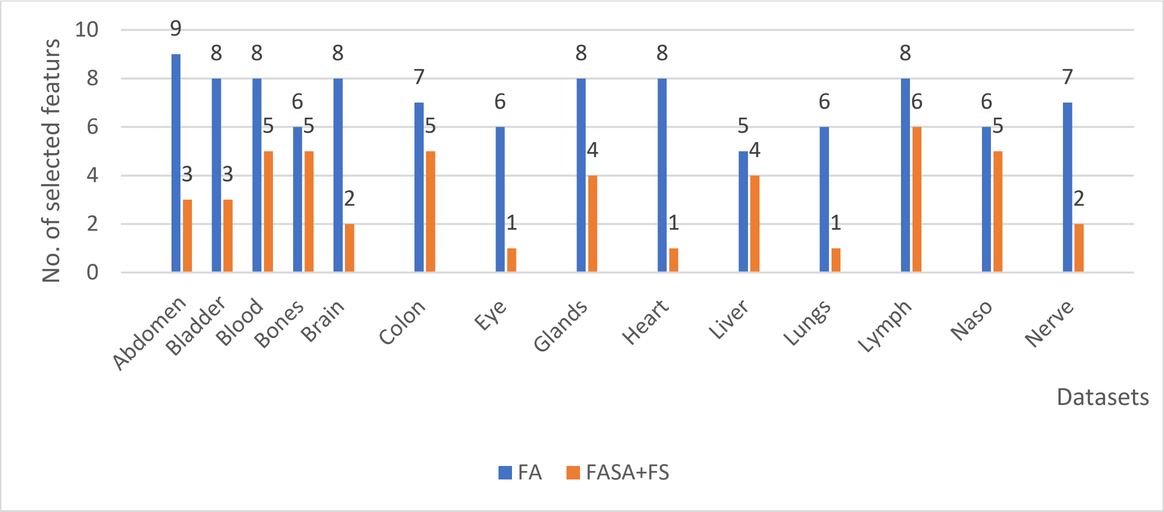


Figure 3: Comparison between FA+FS and FASA+FS number of selected features

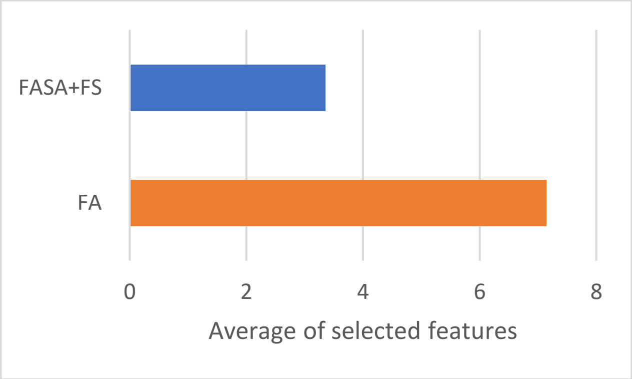


Figure 4: Average no. of selected features by FA+FS and FASA+FS algorithms

6 Conclusions

This paper offered an efficient hybrid strategy for optimizing feature selection based on employing the HRH model. The proposed algorithm detected the minimum and most operative subset of features. FASA+FS combined the FA global search with the SA algorithm. The execution of the proposed approach was evaluated and compared in contrast to FA+FS. Two measures were stated to assess this approach: classification-accuracy and average-feature-selection size. The results showed that the proposed approach achieved an effective balance in both the exploration and exploitation phases.

In all datasets, FASA+FS executed better in standings of classification accuracy than another optimizer, i.e., FA+FS. For future studies, research of the model's presentation on more multifaceted issues is probable.

7 References

Aljarah, I., Habib, M., Faris, H., Al-Madi, N., Heidari, A. A., Mafarja, M., Elaziz, M. A., & Mirjalili, S. (2020). A dynamic locality multi-objective salp swarm algorithm for feature selection. *Computers and Industrial Engineering*, *147*. https://doi.org/10.1016/j.cie.2020.106628

Busetti, F. (2003). Simulated annealing overview. *SAWeb.PDF (geocities.ws)*

Guyon, I., & De, A. M. (2003). An Introduction to Variable and Feature Selection André Elisseeff. In *Journal of Machine Learning Research* (Vol. 3).

Ibrahim, H. T., Mazher, W. J., & Jassim, E. M. (2023). Modified Harris Hawks optimizer for feature selection and support vector machine kernels. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(2). https://doi.org/10.11591/ijeecs.v29.i2.pp942-953

Ibrahim, H. T., Mazher, W. J., Ucan, O. N., & Bayat, O. (2019). A grasshopper optimizer approach for feature selection and optimizing SVM parameters utilizing real biomedical data sets. *Neural Computing and Applications*, *31*(10). https://doi.org/10.1007/s00521-018-3414-4

Kabir, M., Shahjahan, M., & Murase, K. (2013). Ant Colony Optimization Toward Feature Selection. In *Ant Colony Optimization - Techniques and Applications*. https://doi.org/10.5772/51707

Kaya Keles, M., Kilic, U., & Keles, A. E. (2021). Proposed Artificial Bee Colony Algorithm as Feature Selector to Predict the Leadership Perception of Site Managers. *Computer Journal*, *64*(3). https://doi.org/10.1093/comjnl/bxaa163

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

Khushaba, R. N., Al-Ani, A., & Al-Jumaily, A. (2008). Differential Evolution based feature subset selection. *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/icpr.2008.4761255

Mafarja, M. M., & Mirjalili, S. (2017). Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. *Neurocomputing*, *260*. https://doi.org/10.1016/j.neucom.2017.04.053

Ministry of Health. (2017). *Data of Iraqi Cancer Registry for 2010-2012 years*. Iraqi Cancer Board, Ministry of Health.

Mohammadiyan, N. E., & Ghadi, A. (2017). Enhancing the accuracy of firefly algorithm by using the reproduction mechanism. *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 1–6. https://doi.org/10.1109/SNPD.2017.8048202

Oh, I. S., Lee, J. S., & Moon, B. R. (2004). Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(11). https://doi.org/10.1109/TPAMI.2004.105

Pan, X., Xue, L., Lu, Y., & Sun, N. (2019). Hybrid particle swarm optimization with simulated annealing. *Multimedia Tools and Applications*, *78*(21). https://doi.org/10.1007/s11042-018-6602-4

Sun, L., Si, S., Zhao, J., Xu, J., Lin, Y., & Lv, Z. (2023). Feature selection using binary monarch butterfly optimization. *Applied Intelligence*, *53*(1). https://doi.org/10.1007/s10489-022-03554-9

Talbi, E. G. (2002). A taxonomy of hybrid metaheuristics. *Journal of Heuristics*, *8*(5). https://doi.org/10.1023/A:1016540724870

Talbi, E.-G. (2009). *Metaheuristics: From Desing to Implementation.* Published by John Wiley & Sons, Inc., Hoboken, New Jersey and Canada. https://www.wiley.com/en-us/Metaheuristics%3A+From+Design+to+Implementation+-p-9780470278581

Tariq Ibrahim, H., Jalil Mazher, W., Ucan, O. N., & Bayat, O. (2017). Feature Selection using Salp Swarm Algorithm for Real Biomedical Datasets. In *IJCSNS International Journal of Computer Science and Network Security* (Vol. 17, Issue 12).

Yang, X.-She. (2010). *Nature-inspired metaheuristic algorithms*. Luniver Press. https://books.google.com/books/about/Nature_inspired_Metaheuristic_Algorithms.html?id=iVB_ETlh4ogC

Zawbaa, H. M., Emary, E., & Parv, B. (2016). Feature selection based on antlion optimization algorithm. *Proceedings of 2015 IEEE World Conference on Complex Systems, WCCS 2015*. https://doi.org/10.1109/ICoCS.2015.7483317

Zhang, L., Liu, L., Yang, X. S., & Dai, Y. (2016). A novel hybrid firefly algorithm for global optimization. *PLoS ONE*, *11*(9). https://doi.org/10.1371/journal.pone.0163230

ZorarpacI, E., & Özel, S. A. (2016). A hybrid approach of differential evolution and artificial bee colony for feature selection. *Expert Systems with Applications*, *62*. https://doi.org/10.1016/j.eswa.2016.06.004