A Systematic Analysis on the Reproducibility of Results in Bio-Inspired Optimization based Feature Selection Algorithm

A. Anitha
Assistant Professor,
Department of Computer Applications,
The Tamil Nadu Dr. Ambedkar Law University,
Chennai-113, India

ABSTRACT
Reproducibility is a foundation of the scientific method, signifying that when diverse researchers autonomously recreate an experiment by employing the same approaches, they should be reliable and consistently yield the same outcomes. In nature inspired optimization-based feature selection algorithms, irreproducibility may be caused by several factors including the non-convexity nature of the objective, initialization of random values, non-deterministic aspects of training like data shuffling, parallelism, random scheduling, variation in hardware, and round off quantization errors. In this study, the analysis of reproducibility of results in random search bat optimization algorithm for feature selection is conducted for Electrohysterogram (EHG) signals to assess the consistency of the algorithm. The results demonstrated that there was some variability in selected feature sets when the trial process is repeated from 1 to 20 with different bat size. Also, the method is very sensitive to initial parameters in random search process which may require further analysis to improve consistency and robustness. The study's outcomes may underscore the importance of reproducibility in feature selection research, emphasizing that it is crucial for ensuring the robustness and credibility of findings.

Keywords
Bio-Inspired Optimization; Random Search; Reproducibility; Feature Selection; Feature Subsets.

1. INTRODUCTION
Reproducibility stands at the core of the scientific method, indicating that when different researchers independently recreate an experiment or study using the same methods, they should consistently and dependably obtain the same results [1]. This highlights that the findings are not confined to a specific research context but hold validity and applicability across a wider scope. Closely connected to reproducibility is replicability, which entails obtaining results anew by employing the same methods but with fresh data. Replicability serves as confirmation of result reliability, showcasing that they are not influenced by chance variations or sampling errors [2].

Both reproducibility and replicability play a critical role in upholding the quality and credibility of scientific research. They empower researchers to validate their own work and that of their peers, facilitating the accumulation of knowledge. Nonetheless, several challenges exist in the realm of reproducing and replicating research, encompassing issues related to transparency, data accessibility, methodological rigor, and statistical robustness [3].

Nature-inspired optimization algorithms, commonly known as bio-inspired algorithms, draw insights from natural systems and phenomena to tackle complex optimization tasks. These algorithms have seen a growing adoption within the field of feature selection, where the goal is to identify a subset of essential features from a larger pool for use in machine learning models [4]. Bio-inspired techniques provide a valuable strategy for improving random search, a traditional optimization approach, to enhance the effectiveness of feature selection. This collaboration leverages the efficiency of random search while taking advantage of the adaptable nature of bio-inspired algorithms to systematically explore and select feature subsets [5].

Feature selection using random search, while a valuable approach, can encounter several challenges. These challenges can affect the efficiency, effectiveness, and reliability of the feature selection process. In recent years, several researchers have applied feature selection algorithms in the field of physiological signals. These selected features play an important role in the development of classification systems. Several common problems associated with feature selection using random search such as computational complexity, lack of optimality, reproducibility, search space, evaluation metrics, overfitting, etc., [6].

In this study, the analysis of reproducibility of results in random search bio-inspired optimization algorithm for feature selection is conducted for term and preterm EHG signals

2. METHODS
The non-linear signals (Electro hysterogram signals) utilized in this analysis were obtained from the publicly available term preterm EHG database [7]. A total of 76 records were collected, with 38 of them originating from pregnancies beyond the 37th week of gestation. The remaining records pertain to pregnancies that terminated prematurely, occurring before the 37th week of gestation. By employing these signals, the synchro-extracting transform method [8] is applied to produce 2D images that have undergone transformation, from which eighteen distinct features are then extracted.

2.1 Bat Optimization Feature Selection
The Bat algorithm, introduced by Yang in 2010 [9], belongs to the category of nature-inspired intelligence algorithms, drawing inspiration from how bats locate their food sources. This algorithm is particularly useful for addressing intricate nonlinear problems and involves bats emitting a range of signal frequencies spanning

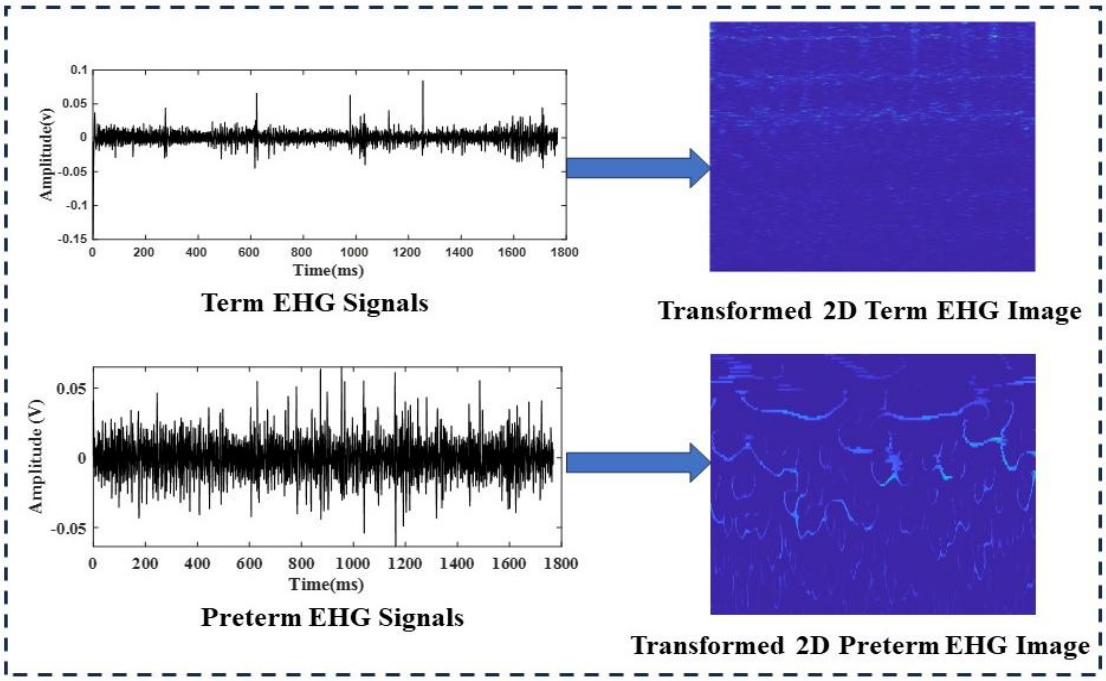

Fig 1: Term and Preterm EHG signals and 2D Transformed Images

from 20KHz to 500 KHz. The algorithm operates by initializing a population of bats, which then update their positions using an echolocation technique based on a similar principle. The fundamental principles of the Bat algorithm, as summarized in references [10], can be stated as follows:

- Bats rely on their echolocation skills to estimate distances and differentiate between potential prey and obstacles.
- When searching for prey, each bat navigates in a random direction with specific attributes such as position, velocity, frequency, wavelength, and loudness. As the search progresses closer to the target, it automatically adjusts the frequency of its emitted pulses and the range of loudness.
- The parameters of pulse emission rate (ranging from 0 to 1), frequency, and loudness are subject to variation.

In this work, reproducibility of results is analyzed using bat optimization feature selection method with various training parameters with number of bat size varying from 10 to 50.

3. RESULTS AND DISCUSSIONS

The eighteen significant time frequency features were extracted from synchro-extracting transformed images of term and preterm EHG signals using Gray level Co-occurrence of Matrix (GLCM). The extracted features are: Autocorrelation (AC), Cluster prominence (CP), cluster Shade (CS), Contrast (C), Correlation (Corr), Difference Variance (DV), Difference Entropy (DE), Dissimilarity (Dis), Energy (E), Entropy (En), Homogeneity (H), Information Measure of Correlation (IMC1 & 2), Inverse Difference (ID), Maximum Probability (MP), Sum Average (SA), Sum Entropy (SE) and Sum of Variance (SV). From these extracted features, significant features are selected using the bat optimization feature selection algorithm for efficient performance for further analysis.

Figure 2 shows the global fitness value of bat optimization-based feature selection method for number of trials from 1 to 20 with different bat sizes. The initial parameters set for the bat optimization algorithm include a population of 10 bats and a maximum of 100 iterations. Using the same initial parameters, the algorithm is executed 20 times. Every execution with different bat size (N), the global fitness value and respective feature sets were noted. It is observed that the global fitness values exhibit a non-linear pattern in the case of random search optimization. The reproducibility of the same feature set is not obtained in the execution process.

Table 1 depicts the different feature subset selected from the overall extracted feature set for every trial. A lack of reproducibility in the selection of feature sets by the algorithm across different trials is observed. The non-reproducibility could be attributed to the random initialization of the algorithm. Small variations in initial conditions or random seeds can lead to divergent solutions. This emphasizes the importance of reporting not only the best result but also the variability in results across trials.

From the results, the irregular pattern and irreproducibility observed in the global fitness value and selected feature sets of bat optimization feature selection method which highly promotes the need of initial parameter considerations. Additionally, the investigation could explore more into understanding the stability, sensitivity to initial parameters and its non-linear dynamics of the algorithm which leading to more consistent and reliable optimization solutions for non-stationary signals.

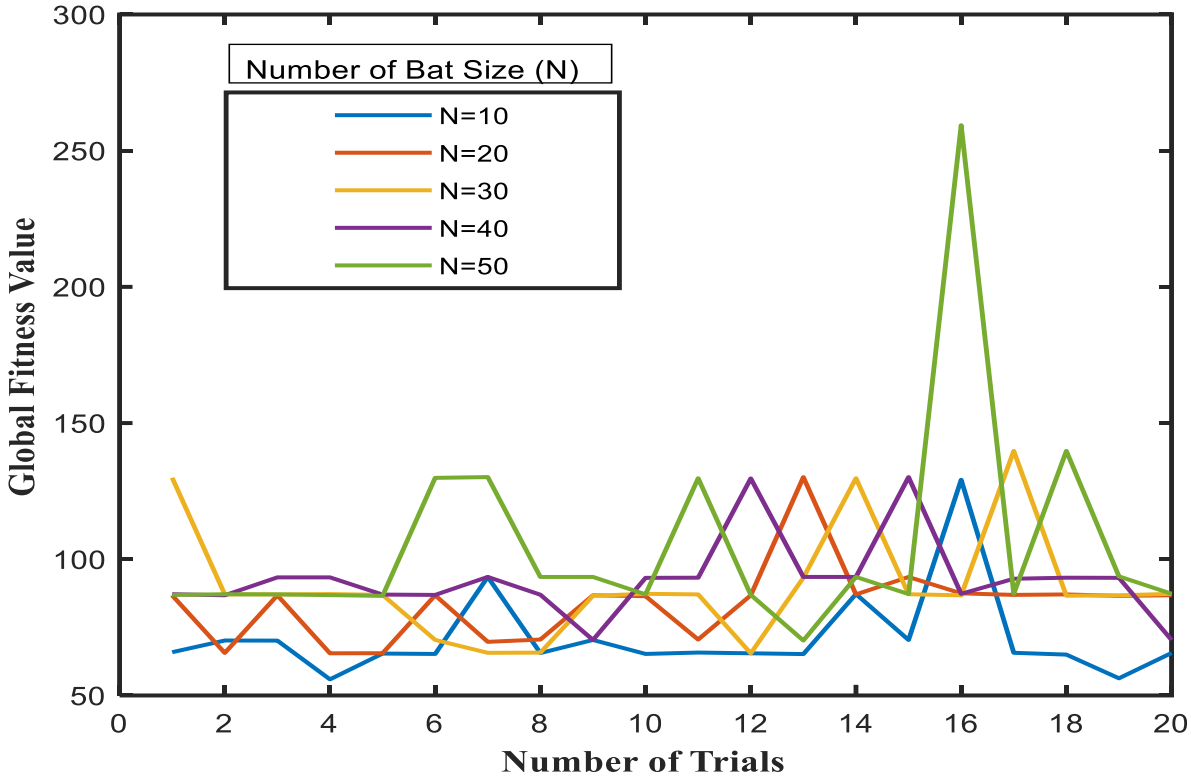

Fig 2: Global Fitness Value of Bat Optimization Feature Selection Method for Number of trials (Bat Size)

Table 1 The selected feature sets for various trial varying from 1 to 20 with different bat size.

Sl. No	Selected Features				
	No of Bat Size (N)				
	N=10	N=20	N=30	N=40	N=50
1	AC, CP, SA, SV	CP, DE, E	CP, IMC2	CP, DE, SA	CP, H, SV
2	CP, CS, Corr, SE	CP, DV, ID, SA	CP, Corr, ID	CP, C, Corr	CP, Dis, SA
3	CP, CS, C, H	CP, DE, E	CP, H, ID	CP, CS, IMC2	CP, ID, MP
4	CP, CS, C, DE, Dis	AC, CP, Dis, ID	CP, SA SE	CP, CS, IMC2	CP, IMC1, SA
5	CP, Corr, IMC1, SA	AC, CP, E, SE	CP, SV, H	AC, CP, En	CP, En, SE
6	AC, CP, DV, DE	CP, DV, ID	AC, CP, CS, SV	CP, Corr, SV	CP, SV
7	CP, H, ID, MP	CP, CS, En, IMC1	CP, E, SA, SE	CP, CS, IMC2	CP, Corr
8	CP, CS, E, H	AC, CP, CS, H	CP, H, SA, SV	AC, CP, Dis	CP, CS, MP
9	AC, CP, IMC1, ID	CP, C, MP	CP, DV, SV	CP, CS, E, H	Cp, CS, ID
10	CP, Corr, IMC2, SA	CP, DE, Dis	AC, CP, H	CP, CS, C	CP, H, ID
11	CP, Dis, SA, SV	AC, CP, CS, Corr	AC, CP, DE, E	CP, CS, SE	CP, SE
12	CP, Dis, E, IMC2	CP, Dis, ID	CP, Corr, H	CP, DE	CP, IMC2, MP
13	CP, En, SA	CP, E	CP, CS, SE	CP, CS, H	CP, CS, ID, SE

14	CP, En, MP	CP, E, H	CP, DE	CP, CS,ID	CP, CS, ID
15	CP, CS, ID, MP	CP, CS, E	CP, Dis, SA	CP, MP	CP, En, SA
16	CP, IMC1	CP, E, SA	CP, Dis, E	AC. CP, Corr	CP
17	CP, H, SA, SE	CP, ID, SV	CP, CS,	CP, CS, IMC1	AC. CP, MP
18	CP, C, En, IMC1	CP, E, H	CP, DE, IMC2	CP, CS, En	CP, CS
19	CP, CS, Corr, DV, ID	CP, En, SE	CP, Dis, E	AC, CP, SV	AC, CP, CS
20	AC, CP, ID, SV	CP, DV, H	AC, CP, ID	CP, DE,En	CP, IMC2, SA

4. CONCLUSION

Reproducibility of findings involves achieving identical results to those of a previous study when conducting an independent investigation, collecting new data, and faithfully following the procedures of the earlier study to the best of one's ability. In the context of nature-inspired optimization, specifically feature selection algorithms, there are various factors that can introduce irreproducibility. These factors include the non-convex nature of the objective function, random initialization, non-deterministic aspects in training like data shuffling, parallelism, random schedules, hardware variations, and round-off quantization errors. This study investigates the reproducibility of results in a random search bat optimization algorithm for feature selection, particularly in the case of term and preterm EHG signals. The analysis aims to assess the consistency of the feature selection in the algorithm search process. The findings revealed some variability in the selected feature sets when the experiment was repeated. Additionally, the method appeared to be sensitive to the initial conditions in the random search process, which may necessitate further examination to enhance its consistency and robustness. These findings underscore the critical role of reproducibility in feature selection research, emphasizing its significance in ensuring the reliability and credibility of results.

5. REFERENCES

[1] Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341), 341ps12-341ps12.

[2] Urkullu, A., Pérez, A., & Calvo, B. (2021). Statistical model for reproducibility in ranking-based feature selection. *Knowledge and Information Systems*, *63*(2), 379-410.

[3] Traverso, A., Wee, L., Dekker, A., & Gillies, R. (2018). Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics*, *102*(4), 1143-1158.

[4] Yang, X. S. (2020). *Nature-inspired optimization algorithms*. Academic Press.

[5] Johnvictor, A. C., Durgamahanthi, V., Pariti Venkata, R. M., & Jethi, N. (2022). Critical review of bio-inspired optimization techniques. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(1), e1528.

[6] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 94.

[7] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

[8] Yu, G., Yu, M., & Xu, C. (2017). Synchroextracting transform. *IEEE Transactions on Industrial Electronics*, *64*(10), 8042-8054.

[9] Gandomi, A. H., Yang, X. S., Alavi, A. H., &Talatahari, S. (2013). Bat algorithm for constrained optimization tasks. *Neural Computing and Applications*, *22*, 1239-1255.

[10] Yang, X. S., & Hossein Gandomi, A. (2012). Bat algorithm: a novel approach for global engineering optimization. *Engineering computations*, *29*(5), 464-483.