TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

Run-Level Comparison of Binary Swarm Optimizers for Gallstone Disease Feature Selection

Jabesh Nehemiah Wijaya*

Faculty of Engineering, Department of Informatics, Universitas Surabaya, Surabaya, Indonesia
Email: jabeshnehemiah@staff.ubaya.ac.id

Abstract–Gallstone disease prediction from non-imaging clinical variables can support early risk assessment, but redundant measurements may reduce model efficiency and complicate interpretation. This study compares five binary swarm optimizers for wrapper feature selection on the UCI Gallstone dataset: Golden Jackal Optimization (bGJO), Egret Swarm Optimization (bESOA), African Vultures Optimization (bAVOA), Tuna Swarm Optimization (bTSO), and FOX Optimization (bFOX). The main contribution of this study is a reproducible run-level statistical comparison of five binary swarm optimizers under an identical wrapper evaluation framework, providing evidence on their accuracy–parsimony and computational trade-offs for gallstone feature selection. The dataset contains 319 complete records, 38 predictors, and an approximately balanced binary outcome. Candidate subsets were evaluated with standardized five-nearest-neighbor classification and stratified 10-fold cross-validation. A weighted objective combined classification error (0.9) and selected-feature proportion (0.1). Each optimizer used 100 epochs, a population of 10, and 10 independent seeded runs. Run-level comparisons used Kruskal–Wallis tests followed by Holm-adjusted Mann–Whitney tests. bGJO achieved the highest mean accuracy (77.56%) with 13.0 features, whereas bAVOA obtained the best mean fitness (0.2270) with only 3.7 features. bTSO selected the fewest features (2.8) but showed greater variability, and bFOX produced the lowest accuracy (67.68%). Accuracy, fitness, feature count, runtime, and memory differed globally among algorithms (all $p<0.001$). The findings demonstrate a clear accuracy–parsimony trade-off: bGJO is preferable when predictive accuracy dominates, while bAVOA offers the strongest overall compromise under the specified objective. External validation and stability analysis of selected variables are required before clinical use.

Keywords: Binary Optimization; Feature Selection; Gallstone Disease; K-Nearest Neighbors; Swarm Intelligence

1. INTRODUCTION

Gallstone disease is a common hepatobiliary disorder in which solid deposits form in the gallbladder or biliary tract. Its clinical importance extends beyond episodic abdominal pain because symptomatic disease can progress to cholecystitis, pancreatitis, biliary obstruction, or other complications. The disease is multifactorial: demographic characteristics, obesity, lipid and glucose metabolism, inflammation, body composition, and genetic or behavioral factors can contribute jointly. Contemporary clinical research therefore increasingly treats risk assessment as a multivariable prediction problem rather than a decision based on one laboratory marker. A study described gallstone disease as closely connected to metabolic dysfunction (Portincasa et al., 2023), while another reported continuing population-level burden (Unalp & Ruhl, 2023). These observations motivate non-imaging tools that can complement—not replace—clinical examination and ultrasonography when rapid screening or risk stratification is needed.

The publicly available Gallstone dataset introduced provides a useful basis for this problem (Esen et al., 2024). It combines demographic information, comorbidity indicators, anthropometric measurements, bioelectrical-impedance variables, liver-related measurements, lipid and glucose markers, renal function, inflammation, hemoglobin, and vitamin D. The cohort comprises 319 individuals, including 161 gallstone cases and 158 healthy controls, and contains no missing observations. Body-composition variables also have clinical precedent: a dataset constructed a gallstone nomogram using sex, body mass index, body-fat percentage, and waist circumference (Lu et al., 2022). Esen et al. (2024) compared nine machine-learning classifiers and reported 85.42% accuracy for gradient boosting. More recent studies have used the same data to examine both predictive performance and model simplification. A study combined rotational forest with bald eagle search, reporting 75.79% accuracy with 17 selected variables (Shrestha et al., 2026), while another reported that an ensemble selection framework enabled random forest to reach 90.62% accuracy with 11 variables (Tamer et al., 2026). These results show that the dataset is sufficiently challenging to expose meaningful differences among classifiers and selection strategies.

Using all available variables is not always desirable. Redundant or weakly relevant measurements can distort distance calculations, increase variance, raise computational cost, and make a clinical model harder to explain. Feature selection retains original measurements while identifying a smaller subset that supports the prediction task. Filter methods rank variables without repeatedly fitting a classifier, embedded methods select variables during model training, and wrapper methods evaluate candidate subsets through a chosen predictive model. Wrapper selection can capture classifier-specific interactions but creates a combinatorial search space of 2^d possible subsets for d predictors. With 38 predictors, exhaustive evaluation would require more than 274 billion subsets. Consequently, approximate optimization is a practical necessity. Recent reviews consistently identify swarm intelligence and other metaheuristics as prominent search mechanisms for wrapper selection because they balance global exploration and local exploitation without requiring differentiability (Barrera-García et al., 2024; Kaur et al., 2023; Rostami et al., 2021).

This study compares five relatively recent nature-inspired optimizers: Golden Jackal Optimization (GJO) (Chopra & Mohsin Ansari, 2022), Egret Swarm Optimization (ESOA) (Chen et al., 2022), African Vultures Optimization (AVOA) (Abdollahzadeh et al., 2021), Tuna Swarm Optimization (TSO) (Xie et al., 2021), and FOX Optimization (Mohammed & Rashid, 2022). Their original continuous forms were proposed for general optimization. Each uses a different

TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

behavioral metaphor and update mechanism, which can produce materially different exploration, convergence, and computational characteristics. Prior work has demonstrated the relevance of binary variants to feature selection. For example, Balakrishnan et al. (2022) evaluated binary AVOA (Balakrishnan et al., 2022), Zhu et al. (2023) enhanced binary GJO (Zhang et al., 2023), and Feda et al. (2024) hybridized FOX with grey-wolf search (Feda et al., 2024). Nevertheless, results across benchmark datasets do not establish which method gives the best accuracy–parsimony compromise for gallstone prediction.

The comparison is designed around a common wrapper evaluator so that differences primarily reflect the optimizers. Every candidate subset is scored by the same standardized five-nearest-neighbor classifier, the same shuffled stratified 10-fold partitions, and the same weighted objective. K-nearest neighbors is appropriate as a transparent wrapper evaluator because irrelevant dimensions directly affect distance-based neighborhoods; feature reduction can therefore have an immediate and measurable effect. The experiment also records runtime and peak traced memory rather than reporting predictive performance alone. This broader evaluation follows recommendations that metaheuristic feature-selection methods should be assessed using accuracy, subset size, convergence, computational cost, and statistical evidence rather than a single best run (Barrera-García et al., 2024; Rostami et al., 2021).

This study evaluates five binary swarm optimizers through an unaggregated, run-level analysis to determine whether their performance differs significantly when every stochastic run is treated as an individual observation. Specifically, the investigation addresses which optimizer achieves the highest cross-validated classification accuracy, which provides the best weighted fitness alongside the strongest feature reduction, and what specific computational trade-offs in runtime and memory consumption accompany each algorithm's performance.

Methodologically, the study contributes a fully reproducible framework applied to a clinically relevant, approximately balanced gallstone dataset, utilizing ten independent seeded runs per optimizer to capture genuine stochastic variability rather than relying on a single best run. To ensure analytical rigor, all comparisons are subjected to strict statistical control using global significance tests, multiplicity-adjusted post-hoc pairwise tests, and effect size calculations.

Finally, the paper delivers a multi-dimensional performance evaluation that balances cross-validated accuracy, weighted fitness, selected feature counts, computational resource costs, and overall convergence behavior. These insights provide empirical guidance for selecting an optimal feature selection strategy in subsequent gallstone prediction and validation research, serving as a practical decision framework rather than proposing a standalone diagnostic device.

2. RESEARCH METHODOLOGY

2.1 Research Design and Dataset

This computational experiment used only the local copy of the UCI Gallstone dataset (Kelly et al., n.d.). The source cohort was collected at the Internal Medicine Outpatient Clinic of Ankara VM Medical Park Hospital between June 2022 and June 2023 under ethics approval E2-23-4632. After excluding individuals with previous gallbladder surgery, 319 records remained. The target, Gallstone Status, was binary, with 161 positive and 158 negative records. Thirty-eight predictors described age, sex, comorbidity, coronary artery disease, hypothyroidism, hyperlipidemia, diabetes, height, weight, body mass index, body-water compartments, body-fat and lean-mass measurements, visceral measures, hepatic fat accumulation, glucose, lipid profile, liver enzymes, alkaline phosphatase, creatinine, glomerular filtration rate, C-reactive protein, hemoglobin, and vitamin D. The workbook contained no missing values; nevertheless, complete-case checking was retained in the pipeline as a defensive validation step. The research procedure consisted of six main stages: dataset preparation, preprocessing, wrapper construction, binary swarm optimization, repeated experimental evaluation, and statistical comparison. First, the Gallstone dataset was validated and separated into predictors and target variables. Second, candidate feature subsets generated by each optimizer were evaluated using standardized KNN with stratified 10-fold cross-validation. Third, classification error and feature-subset size were combined through the weighted fitness function. Fourth, each binary optimizer was executed in ten independently seeded runs under an identical computational budget. Fifth, accuracy, fitness, selected-feature count, runtime, memory consumption, and convergence histories were collected. Finally, the run-level results were statistically compared using Kruskal–Wallis and Holm-adjusted Mann–Whitney tests.

All computational workflows were implemented in Python, leveraging a robust analytical stack that included NumPy, pandas, scikit-learn, SciPy, joblib, and the MEALPY optimization framework. To ensure data integrity during ingestion, the dataset file—which utilizes Strict Office Open XML namespaces—was processed through a custom loader that normalized these namespace identifiers in memory without altering the underlying source workbook.

Data preprocessing was structured to guarantee consistent variable formatting and prevent systemic downstream errors. The target column was explicitly separated by name rather than relying on its positional index, safeguarding against scenarios where non-predictive variables (such as ID markers in the first column) might be mistakenly treated as feature predictors. Additionally, any categorical predictors present in the dataset were one-hot encoded, and the complete feature matrix was cast to 32-bit floating-point precision to optimize computational performance during optimization. The overall structure of this experimental pipeline, from raw data loading and preprocessing to feature selection and evaluation, is systematically outlined in the methodology flowchart presented in Figure 1.

TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

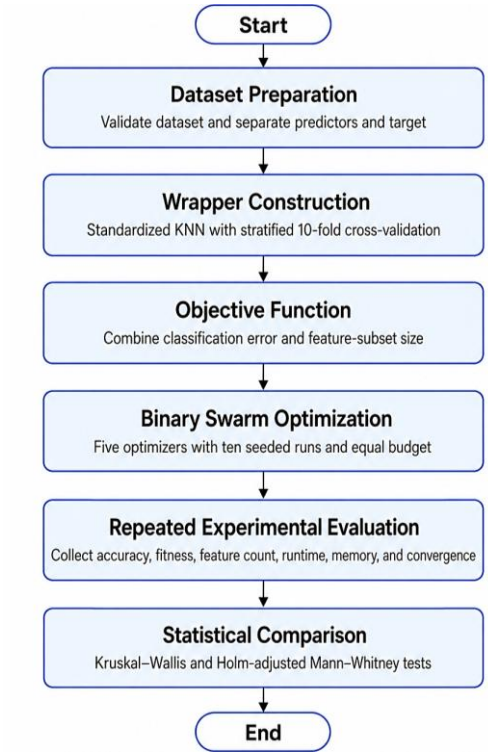

Figure 1. Research flowchart

2.2 Wrapper Evaluation and Objective Function

Candidate feature subsets were evaluated with a pipeline containing standard-score normalization followed by a K-nearest-neighbor classifier with k=5. Scaling was fitted independently inside each training fold, preventing information from the validation fold from influencing normalization. Performance was estimated using shuffled stratified 10-fold cross-validation with random state 42 and classification accuracy as the score. Stratification preserved the nearly balanced class proportions in every fold.

Each optimizer represented a solution as 38 continuous values bounded from −8 to 8. The logistic transfer function $T(x) = 1/(1 + e\text{^}(-x))$ converted each component, and a predictor was retained when $T(x) \geq 0.5$. An all-zero subset received the worst objective vector. For a nonempty subset S, the two objectives were classification error and the proportion of selected predictors. They are then combined as $J(S) = 0.9[1 - Accuracy(S)] + 0.1[|S|/38]$, where lower fitness is better. The 0.9 weight made predictive accuracy the primary aim, while the 0.1 term rewarded parsimonious subsets and resolved near-ties. This structure resembles recent wrapper formulations that optimize classifier error and subset size jointly (Balakrishnan et al., 2022; Feda et al., 2024; Xu et al., 2021).

2.3 Optimizers and Experimental Protocol

The evaluated methods were binary implementations of GJO, ESOA, AVOA, TSO, and FOX, denoted bGJO, bESOA, bAVOA, bTSO, and bFOX. GJO models cooperative hunting by a golden-jackal pair; ESOA combines sit-and-wait, aggressive, and discriminant search strategies; AVOA models hunger-driven movement and competition among vultures; TSO uses spiral and parabolic foraging; and FOX models prey localization and jumping behavior. Although the metaphors differ, every method received the same representation, bounds, transfer rule, evaluator, population size, and stopping budget. This controls several common sources of unfairness in metaheuristic comparisons.

Each algorithm used 100 epochs and a population of 10. Ten runs were performed per algorithm, producing 50 run-level results. Run r used seed $42 + r - 1$, so the run sequence was reproducible while retaining stochastic diversity. For each run, the final weighted fitness, cross-validated accuracy, selected-feature count, optimization time, peak Python-traced memory, and global-best fitness history were stored. Parallel execution accelerated repeated runs; consequently, absolute runtime should be interpreted for this software and hardware environment rather than as a universal benchmark.

2.4 Statistical Analysis

For each metric and algorithm, the analysis calculated count, mean, sample standard deviation, median, minimum, and maximum across the ten runs. Because sample sizes were small and stochastic optimizer outcomes need not be normally distributed, omnibus comparisons used the Kruskal–Wallis rank test. When a global test was significant at $\alpha = 0.05$, all ten algorithm pairs were examined with two-sided Mann–Whitney U tests. Holm adjustment controlled the family-wise error rate separately within each metric. Rank-biserial correlation was reported as an effect-size measure; its sign indicates whether values from the first named algorithm tended to be larger or smaller than values from the second. For accuracy,

TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

a positive sign favors the first algorithm, whereas for fitness, runtime, memory, and feature count the practical preference generally lies with smaller values. The statistical calculations used SciPy (Virtanen et al., 2020). Table 1 provides the configurations used in this research.

Table 1. Dataset and experimental configuration

Component	Configuration
Dataset	UCI Gallstone; 319 records; 161 positive and 158 negative
Predictors	38 demographic, comorbidity, body-composition, and laboratory variables
Classifier	StandardScaler + KNN (k=5)
Validation	Shuffled stratified 10-fold CV; random state 42
Optimizers	bGJO, bESOA, bAVOA, bTSO, and bFOX
Search budget	100 epochs; population 10; bounds [−8, 8]
Repeated runs	10 per algorithm; seeds 42–51
Objective	0.9 × classification error + 0.1 × selected-feature ratio

3. RESULTS AND DISCUSSION

3.1 Descriptive Performance

Table 2 summarizes all ten runs per algorithm. bGJO produced the highest mean accuracy, 0.7756, with a sample standard deviation of 0.0112 and a range of 0.7527–0.7902. Its mean subset contained 13.0 predictors, corresponding to a 65.8% reduction from the original 38. bESOA was second in mean accuracy at 0.7688, selected 13.6 predictors, and displayed a similar accuracy spread. Its major disadvantage was computational: mean runtime reached 296.37 seconds, almost three times the bGJO mean and more than three times those of bTSO and bFOX.

bAVOA produced the lowest mean weighted fitness, 0.2270, even though its mean accuracy of 0.7586 was lower than those of bGJO and bESOA. This result follows directly from the objective design: bAVOA retained only 3.7 predictors on average, a 90.3% reduction, and therefore gained a substantial parsimony advantage. The median subset was only two predictors, although the range of 1–13 shows that solutions varied considerably among runs. Its median fitness of 0.2209 was also the lowest observed algorithm-level median.

bTSO was the most aggressive reducer, selecting 2.8 predictors on average and only 1.5 at the median. Its mean accuracy was 0.7486. However, variability was visibly larger than for bGJO, bESOA, and bAVOA: accuracy ranged from 0.6804 to 0.7589 and fitness from 0.2223 to 0.2981. The large difference between mean fitness (0.2337) and median fitness (0.2224) indicates that at least one poor run pulled the mean upward. Thus, bTSO sometimes found highly compact competitive subsets but was less reliable across initializations.

bFOX was dominated on the principal predictive outcomes. It selected 18.4 predictors on average—more than any competitor, yet achieved the lowest mean accuracy (0.6768) and highest mean fitness (0.3393). Its only clear advantages were mean peak traced memory (2.8374 MB, the lowest) and mean runtime (86.05 seconds, narrowly below bTSO). The result illustrates why computational economy cannot be interpreted apart from solution quality: bFOX was inexpensive in the measured environment but did not convert that economy into a useful accuracy–parsimony compromise.

Table 2. Run-level performance across ten runs

Algorithm	Accuracy mean±SD	Fitness mean±SD	Features mean±SD	Time (s) mean±SD	Memory (MB) mean±SD
bGJO	0.7756±0.0112	0.2362±0.0082	13.0±2.16	107.68±8.21	3.4968±0.1105
bESOA	0.7688±0.0122	0.2438±0.0116	13.6±2.99	296.37±7.60	3.8792±0.1744
bAVOA	0.7586±0.0126	0.2270±0.0162	3.7±3.65	102.26±1.86	4.0514±0.1291
bTSO	0.7486±0.0241	0.2337±0.0255	2.8±3.71	87.48±3.85	4.6874±0.0911
bFOX	0.6768±0.0139	0.3393±0.0126	18.4±2.76	86.05±10.92	2.8374±0.2078

3.2 Run-Level Statistical Comparisons

The omnibus results in Table 3 reject equality among algorithms for every recorded metric. Accuracy differed with H=34.6415 ($p = 1.0 \times 10\text{^} - 6$), and fitness differed with H=32.2966 ($p = 2.0 \times 10\text{^} - 6$). Selected-feature count, runtime, and peak memory also produced highly significant global differences (all $p < 10\text{^} - 6$ after rounding in the notebook output). These tests use all 50 run-level observations; they therefore establish that the distributions, not merely the displayed means, differed across methods.

Holm-adjusted accuracy comparisons shown in Table 4 refine this conclusion. bGJO significantly exceeded bAVOA ($adjusted\ p = 0.017676$; $rank - biserial\ r = 0.78$), bTSO ($p = 0.009279$; $r = 0.84$), and bFOX ($p = 0.001756$; $r = 1.00$), but not bESOA ($p = 0.173456$). The bGJO–bFOX effect was complete in the observed runs: every bGJO result ranked above every bFOX result. bESOA also exceeded bFOX ($p = 0.001756$; $r = 1.00$), as did bAVOA and bTSO. Differences among bESOA, bAVOA, and bTSO did not survive Holm correction; the bESOA–

TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

bAVOA and bESOA–bTSO adjusted p-values were both 0.052665, just above the prespecified threshold. Reporting these nonsignificant comparisons is important because the ordering of means alone might otherwise be overstated.

Fitness comparisons tell a different story. Although bAVOA had the lowest mean fitness, none of the pairwise differences among bGJO, bESOA, bAVOA, and bTSO remained significant after adjustment. Every one of these four algorithms significantly outperformed bFOX, with adjusted p-values of 0.001593 and absolute rank-biserial effects of 1.00. The absence of significant differences among the four leading methods reflects both the weighted objective and run variability. It means that choosing among them should rely on the desired operating point—maximum accuracy, maximum reduction, or stability—not on a claim that one uniformly minimizes fitness.

Feature-count comparisons separated the methods into interpretable groups. bGJO and bESOA did not differ ($adjusted\ p = 0.568120$), and bAVOA and bTSO did not differ ($p = 0.263917$). All contrasts between these high-feature and low-feature groups were significant. bFOX also differed significantly from every method because it retained the largest subsets. Hence, the selection behavior was not random noise around a common feature count; the optimizers adopted systematically different positions on the accuracy–parsimony surface.

Runtime and memory results further qualify the ranking. bESOA was significantly slower than every competitor. bGJO and bAVOA were not distinguishable in runtime after adjustment, while bAVOA, bTSO, and bFOX contained some nonsignificant pairwise runtime differences. Peak traced memory differed significantly for all pairs, with bFOX the lowest and bTSO the highest. These memory values represent allocations visible to Python's tracing mechanism and do not include every native allocation by NumPy or worker processes, so they are best treated as relative implementation indicators.

Table 3. Kruskal–Wallis global comparisons

Metric	H statistic	p value	Decision at α=0.05
Accuracy	34.6415	0.000001	Significant
Fitness	32.2966	0.000002	Significant
Selected features	37.9368	<0.000001	Significant
Runtime	38.2993	<0.000001	Significant
Peak memory	45.6696	<0.000001	Significant

Table 4. Holm-adjusted pairwise accuracy comparisons

Comparison	Rank-biserial	Adjusted p	Significant
bGJO vs bESOA	0.37	0.173456	No
bGJO vs bAVOA	0.78	0.017676	Yes
bGJO vs bTSO	0.84	0.009279	Yes
bGJO vs bFOX	1.00	0.001756	Yes
bESOA vs bAVOA	0.60	0.052665	No
bESOA vs bTSO	0.66	0.052665	No
bESOA vs bFOX	1.00	0.001756	Yes
bAVOA vs bTSO	0.65	0.052665	No
bAVOA vs bFOX	1.00	0.001756	Yes
bTSO vs bFOX	0.90	0.004787	Yes

3.3 Convergence and Accuracy–Parsimony Trade-off

Figure 2 plots average best fitness by epoch. The curves summarize search dynamics over ten runs rather than displaying a single favorable trajectory. The figure should be read together with final distributions: rapid apparent convergence is desirable only if it reaches a competitive terminal value. bAVOA and bTSO achieved strong compact solutions, whereas bGJO maintained the best predictive accuracy. bESOA's longer runtime did not yield a statistically significant accuracy improvement over bGJO or a significant fitness advantage over the other leading algorithms. bFOX converged to a visibly weaker region consistent with its final fitness statistics.

The weighted objective explains why the accuracy winner and fitness winner differ. Relative to bGJO, bAVOA sacrificed approximately 1.70 percentage points of mean accuracy but removed an additional 9.3 predictors on average. bTSO sacrificed 2.70 points while removing 10.2 additional predictors. If the study objective is screening performance under the present KNN evaluator, bGJO is the defensible choice because it achieved the largest mean accuracy and significantly exceeded three competitors. If measurement burden and model simplicity are important, bAVOA offers the strongest objective-defined compromise and retained roughly one tenth of the original predictors. bTSO may be attractive under extremely restrictive measurement budgets, but its wider range calls for more runs or stability controls.

The distinction has practical implications. A reduced feature panel could lower data-acquisition burden only if the chosen variables are stable, clinically obtainable, and validated. The current notebook recorded the number of retained predictors but did not persist the identity of the selected subset from each run. It is therefore not possible to claim that bAVOA consistently chose a specific laboratory or body-composition marker. That interpretation must await a revised experiment that stores the binary mask and selection frequency. This limitation is especially important because Esen et al. (2024) highlighted vitamin D, C-reactive protein, total body water, and fat-free mass, while Shrestha et al. (2026)

TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

emphasized overlapping but not identical variables including C-reactive protein, vitamin D, obesity, hemoglobin, and bone mass. Agreement in subset size is not evidence of agreement in clinical content.

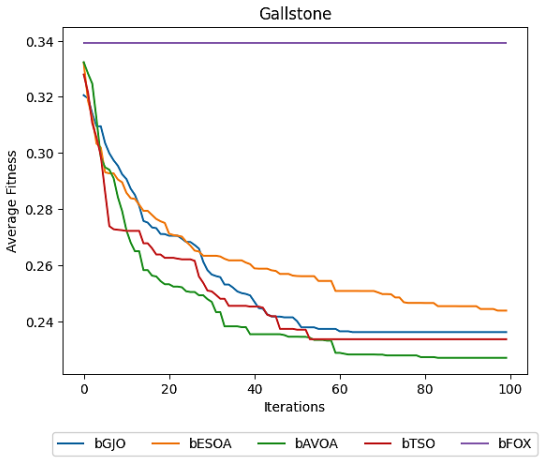

Figure 2. Mean global-best fitness convergence over 100 epochs

3.4 Comparison

The best accuracy in the present comparison, 77.56% for bGJO, is lower than the 85.42% gradient-boosting result reported by Esen et al. (2024). This difference should not be interpreted as a failure of feature selection because the predictive learners and experimental goals differ. The present study deliberately fixed a simple KNN classifier and optimized a two-component objective, whereas gradient boosting learns nonlinear additive decision structures and was evaluated as the final classifier. The current result is closer to the 75.79% obtained by Shrestha et al. (2026) after bald-eagle feature optimization of rotational forest, although their model retained 17 variables and their cross-validation framework differs. Tamer et al. (2026) subsequently reported 90.62% with 11 features using an ensemble selection scheme and random forest, demonstrating the potential benefit of evaluating selected panels across stronger classifiers.

The comparison with prior binary metaheuristics is also consistent with a broader pattern: no optimizer dominates every dataset and metric. Binary AVOA variants have produced strong classification and reduction results on benchmark collections (Balakrishnan et al., 2022), and improved binary GJO has been designed specifically to prevent premature convergence (Zhang et al., 2023). In biomedical classification, a study likewise demonstrated that dynamic search modifications can materially change feature-selection performance (Houssein et al., 2023). Feda et al. (2024) showed that a hybrid FOX variant can outperform the original search on many feature-selection datasets. The poor bFOX outcome here therefore applies to the implemented original binary adaptation and this dataset; it should not be generalized to all FOX hybrids. Likewise, the efficiency of bAVOA in this experiment does not prove universal superiority. Reviews of swarm feature selection emphasize that representation, binarization, classifier choice, objective weights, and dataset characteristics can change rankings substantially (Barrera-García et al., 2024; Rostami et al., 2021).

From a methodological perspective, the run-level analysis improves on reporting only a best solution. Best-run reporting is optimistically biased because stochastic search receives multiple opportunities to produce an extreme value. Means and medians quantify typical performance, ranges reveal instability, and rank tests evaluate whether observed distributions are plausibly interchangeable. The bTSO findings are a clear example: its best fitness (0.2223) was competitive, but its worst fitness (0.2981) and broad accuracy range reveal risk that a best-only table would conceal. Conversely, bGJO's relatively narrow accuracy distribution supports its use when repeatability matters.

The statistical design nevertheless has boundaries. The Mann–Whitney comparisons treat runs as independent stochastic realizations on one fixed dataset. The resulting p-values characterize optimizer behavior for this dataset and pipeline; they do not establish general superiority across clinical populations or benchmark problems. A multi-dataset study would support blocked tests across datasets, while a clinical validation study would require patient-level uncertainty, discrimination and calibration measures, and an external cohort. The global tests were also performed for five outcomes; Holm control was applied within each pairwise family but not across the five omnibus endpoints. Accuracy and fitness should therefore be regarded as the primary outcomes, with runtime and memory as supporting evidence.

A further concern is selection bias within cross-validation. The wrapper repeatedly queries the same ten folds during optimization, so the reported cross-validated accuracy participates in selecting the feature subset. This can yield optimistic performance even though scaling is correctly confined to each fold. Nested cross-validation would place the complete feature-selection search inside an outer training loop and evaluate the chosen subset on unseen outer folds. External validation would be stronger still. These steps are essential before comparing the reported accuracy directly with a clinical diagnostic threshold.

Clinical interpretation also requires attention to the asymmetric consequences of error. Accuracy assigns equal cost to a false-negative gallstone prediction and a false-positive referral, but missed disease may delay follow-up while false-positive predictions may prompt unnecessary testing. The nearly balanced dataset makes accuracy understandable


TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

for algorithm comparison, yet it does not remove this clinical asymmetry. A subsequent evaluation should specify the intended screening pathway, select an operating threshold, and report class-specific confidence intervals. Calibration is equally important: an accurate rank ordering does not guarantee that a predicted probability represents observed risk. These considerations reinforce that the current ranking concerns feature-search behavior under a controlled evaluator, not readiness for autonomous diagnosis.

The results indicate that optimizer selection should not be based on a single performance metric. bGJO achieved the highest mean classification accuracy, suggesting that its search behavior was more effective in preserving informative predictors that supported KNN discrimination. However, this advantage was obtained with a larger feature subset than those selected by bAVOA and bTSO. In contrast, bAVOA achieved the best mean weighted fitness because the objective function explicitly rewarded both predictive performance and feature reduction. This explains why bAVOA could outperform bGJO in fitness despite having slightly lower classification accuracy.

The findings also highlight an important accuracy–parsimony trade-off. A larger feature subset may preserve more predictive information, but it increases model complexity and potential data-acquisition burden. Conversely, a very small subset can improve simplicity and efficiency but may reduce predictive stability if relevant variables are removed. In this study, bTSO demonstrated the strongest feature reduction, although its larger variability across runs suggests that aggressive reduction may also produce less consistent solutions.

Therefore, no single optimizer can be considered universally superior. bGJO is more appropriate when predictive accuracy is the primary objective, whereas bAVOA is more suitable when a balance between accuracy and feature reduction is required. bTSO may be considered when minimizing the number of retained variables is the main priority. These results reinforce the importance of evaluating stochastic feature-selection algorithms using multiple criteria and repeated runs rather than selecting an optimizer from a single best solution.

4. CONCLUSION

This study compared five binary swarm optimizers for wrapper feature selection on 319 complete gallstone records with 38 predictors. Under a common standardized five-nearest-neighbor evaluator, stratified 10-fold cross-validation, 100 epochs, a population of 10, and ten seeded runs, bGJO achieved the highest mean accuracy of 77.56% while retaining 13.0 predictors. bAVOA achieved the best mean weighted fitness of 0.2270 with only 3.7 predictors, and bTSO produced the smallest mean subset of 2.8 predictors but with greater variability. bFOX was fastest and used the least traced memory, yet it had the lowest accuracy and largest selected subset, making it unsuitable under the present objective. Run-level Kruskal–Wallis tests found significant algorithm differences for all five outcomes, while Holm-adjusted post-hoc results showed that bGJO's accuracy advantage was significant over bAVOA, bTSO, and bFOX but not bESOA. The results answer the research questions by identifying bGJO as the accuracy-oriented choice and bAVOA as the most favorable accuracy–parsimony compromise. The study remains a computational comparison rather than a validated clinical model. Its main limitations are the single modest-sized cohort, reuse of cross-validation folds during wrapper search, absence of persisted feature identities, and hardware-dependent resource measurements. Future work should use nested and external validation, record selection stability, test consensus subsets with multiple calibrated classifiers, and report sensitivity, specificity, area under the curve, and clinical utility before deployment is considered.

REFERENCES

Abdollahzadeh, B., Gharehchopogh, F. S., & Mirjalili, S. (2021). African vultures optimization algorithm: A new nature-inspired metaheuristic algorithm for global optimization problems. *Computers & Industrial Engineering*, *158*, 107408. https://doi.org/10.1016/j.cie.2021.107408

Balakrishnan, K., Dhanalakshmi, R., & Seetharaman, G. (2022). S-shaped and V-shaped binary African vulture optimization algorithm for feature selection. *Expert Systems*, *39*(10), e13079. https://doi.org/10.1111/exsy.13079

Barrera-García, J., Cisternas-Caneo, F., Crawford, B., Gómez Sánchez, M., & Soto, R. (2024). Feature Selection Problem and Metaheuristics: A Systematic Literature Review about Its Formulation, Evaluation and Applications. *Biomimetics*, *9*(1), 9. https://doi.org/10.3390/biomimetics9010009

Chen, Z., Francis, A., Li, S., Liao, B., Xiao, D., Ha, T. T., Li, J., Ding, L., & Cao, X. (2022). Egret Swarm Optimization Algorithm: An Evolutionary Computation Approach for Model Free Optimization. *Biomimetics*, *7*(4), 144. https://doi.org/10.3390/biomimetics7040144

Chopra, N., & Mohsin Ansari, M. (2022). Golden jackal optimization: A novel nature-inspired optimizer for engineering applications. *Expert Systems with Applications*, *198*, 116924. https://doi.org/10.1016/j.eswa.2022.116924

Esen, İ., Arslan, H., Aktürk Esen, S., Gülşen, M., Kültekin, N., & Özdemir, O. (2024). Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data. *Medicine*, *103*(8), e37258. https://doi.org/10.1097/MD.0000000000037258

Feda, A. K., Adegboye, M., Adegboye, O. R., Agyekum, E. B., Fendzi Mbasso, W., & Kamel, S. (2024). S-shaped grey wolf optimizer-based FOX algorithm for feature selection. *Heliyon*, *10*(2), e24192. https://doi.org/10.1016/j.heliyon.2024.e24192


TIN: Terapan Informatika Nusantara
Vol 7, No 3, August 2026, page 1587-1594
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i3.10978

Houssein, E. H., Samee, N. A., Mahmoud, N. F., & Hussain, K. (2023). Dynamic Coati Optimization Algorithm for Biomedical Classification Tasks. *Computers in Biology and Medicine*, *164*, 107237. https://doi.org/10.1016/j.compbiomed.2023.107237

Kaur, S., Kumar, Y., Koul, A., & Kumar Kamboj, S. (2023). A Systematic Review on Metaheuristic Optimization Techniques for Feature Selections in Disease Diagnosis: Open Issues and Challenges. *Archives of Computational Methods in Engineering*, *30*(3), 1863–1895. https://doi.org/10.1007/s11831-022-09853-1

Kelly, M., Longjohn, R., & Nottingham, K. (n.d.). *The UCI Machine Learning Repository*. Retrieved http://archive.ics.uci.edu

Lu, J., Tong, G., Hu, X., Guo, R., & Wang, S. (2022). Construction and Evaluation of a Nomogram to Predict Gallstone Disease Based on Body Composition. *International Journal of General Medicine*, *15*, 5947–5956. https://doi.org/10.2147/IJGM.S367642

Mohammed, H., & Rashid, T. (2022). FOX: A FOX-inspired optimization algorithm. *Applied Intelligence*, *53*(1), 1030–1050. https://doi.org/10.1007/s10489-022-03533-0

Portincasa, P., Di Ciaula, A., Bonfrate, L., Stella, A., Garruti, G., & Lamont, J. T. (2023). Metabolic dysfunction-associated gallstone disease: Expecting more from critical care manifestations. *Internal and Emergency Medicine*, *18*(7), 1897–1918. https://doi.org/10.1007/s11739-023-03355-z

Rostami, M., Berahmand, K., Nasiri, E., & Forouzandeh, S. (2021). Review of swarm intelligence-based feature selection methods. *Engineering Applications of Artificial Intelligence*, *100*, 104210. https://doi.org/10.1016/j.engappai.2021.104210

Shrestha, K., Sarker, P., Tiang, J.-J., & Nahid, A.-A. (2026). Metaheuristic-based gallstone classification using rotational forest explained with SHAP. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1727559

Tamer, C. Ç., Arslan, H., & Çağlıkantar, T. (2026). A novel method for predicting gallstones based on ensemble feature selection method. *Scientific Reports*. https://doi.org/10.1038/s41598-026-53394-7

Unalp, A., & Ruhl, C. (2023). Increasing gallstone disease prevalence and associations with gallbladder and biliary tract mortality in the United States. *Hepatology*, *Publish Ahead of Print*. https://doi.org/10.1097/HEP.0000000000000264

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Xie, L., Han, T., Zhou, H., Zhang, Z.-R., Han, B., & Tang, A. (2021). Tuna Swarm Optimization: A Novel Swarm-Based Metaheuristic Algorithm for Global Optimization. *Computational Intelligence and Neuroscience*, *2021*(1), 9210050. https://doi.org/10.1155/2021/9210050

Xu, Y., Huang, H., Heidari, A. A., Gui, W., Ye, X., Chen, Y., Chen, H., & Pan, Z. (2021). *MFeature*: Towards high performance evolutionary tools for feature selection. *Expert Systems with Applications*, *186*, 115655. https://doi.org/10.1016/j.eswa.2021.115655

Zhang, K., Liu, Y., Mei, F., Sun, G., & Jin, J. (2023). IBGJO: Improved Binary Golden Jackal Optimization with Chaotic Tent Map and Cosine Similarity for Feature Selection. *Entropy*, *25*(8), 1128. https://doi.org/10.3390/e25081128