Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

SYNERGISTIC INTEGRATION OF WHALE, PARTICLE SWARM, AND FIREFLY BASED TRIBIOINSPIRED MULTI-OBJECTIVE HYBRID OPTIMIZATION PARADIGMS FOR ENHANCED BIG DATA CLASSIFICATION PERFORMANCE

ABHIMANYU DUTONDE*
Research Scholar, PGTD Computer Science and Electronics Department, RTMNU, Nagpur, India.
Tulsiramji Gaikwad-Patil College of Engineering and Technology, Nagpur.
*Corresponding Author Email: abhimanyudutonde@gmail.com

SHRIKANT SONEKAR
Associate Professor, J D College of Engineering and Management, Nagpur, India.
Email: srikantsonekar@gmail.com

Abstract

This paper presents a novel hybrid tribioinspired model for feature selection that employs three algorithms for feature selection: Whale Optimization for Interclass Variance Maximization, Particle Swarm Optimization for Intraclass Variance Minimization, and Firefly Optimization for Best Weights Selection. The WOICVM algorithm is used as this contains an excellent exploration-exploitation balance for maximizing the interclass variance to induce significant separability among classes. PSOICVM is highly efficient in large-scale optimization, minimizes intraclass variance, and improves cohesiveness within each class. Finally, the Firefly Algorithm optimally combines WOICVM's and PSOICVM's strengths by determining the best weighting scheme and balancing interclass and intraclass variances. This multiobjective approach enhances feature selection efficiency by leveraging the complementary advantages of the three algorithms. Tentative numerical results depict a 15% increase in inter-class variance with WOICVM, a 12% reduction in intraclass variance with PSOICVM, and a 20% improvement in overall feature selection efficiency through FOBWS. This thereby shows a 10% enhancement in classification accuracy in high-dimensional environments, showing the efficiency of the proposed model over conventional methods. It fills critical gaps in existing methods by offering a hybrid method as a strong tool for applying big data to improve classification performance.

Keywords: Big Data, Feature Selection, Whale Optimization, Particle Swarm Optimization, Firefly Algorithm, Process.

1. INTRODUCTION

Big data, which explodes exponentially in different fields of health care, finance, and social networks, has increased the need for advanced analytical tools to process vast volumes of high-dimensional datasets and analyze them efficiently. More specifically, feature selection has emerged as one of the critical tasks at the data preprocessing stage to reduce dimensionality while retaining the most informative features. The central challenge in feature selection lies with the balance between interclass variance, its ability to distinguish one class from another, and intraclass variance, its ability to maintain cohesion within the same class. Generally speaking [1, 2, 3], traditional feature selection methods fail in this task since they rely on simple heuristics or linear models that cannot

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

grasp complexity and scale effectively; the features can be involved in relationships of high dimensionality as is often found in most datasets and samples. Conventional techniques like PCA, mutual information-based selection and filter-based approaches relate very much either to a reduction of the dimensionality of the search space or to an improvement in the performance of the pattern classification, but they can hardly do both. Moreover, such approaches assume a linear relationship between the features and neglect the nonlinear interactions that are, in fact, quite common for high-dimensional data samples. This would naturally result in suboptimal performance, particularly when complex class structures exist or numerous irrelevant and redundant features are present in the given datasets. The inability of such traditional approaches to address the dynamic and complex nature of big data environments has made it a pressing need to develop novel, more efficient optimization techniques that can handle both multidimensional aspects of data together while ensuring the best possible accuracy in classification. With the challenge above [4, 5, 6], much focus has been put into bio-inspired optimization algorithms that could explore large search spaces efficiently and discover at least near-optimal solutions within a reasonable time. Some of the most promising ones are PSO, WOA, and FA. These algorithms are particularly effective in the feature selection task scenario with big data environments because they balance exploration and exploitation, two main factors contributing to the optimization solution in high-dimensional spaces. This paper proposes a new integrated model for feature selection that uses the strengths of such bio-inspired algorithms. The Whale Optimization for Interclass Variance Maximization specifically looks into maximizing the class separation through the rates of variance levels. This is in harmony with particle swarm optimization for intraclass variance minimization, which minimizes the variance in each class and continues to polish up homogeneity among features of the same class. Finally, the FOBWS uses the Firefly optimization algorithm for the best weight selection to seek the optimal weighting combination of these two methods. Therefore, it balances the strengths of these two methods and makes feature selection more efficient. The integrated approach gives a much more robust answer to feature selection in high-dimensional big-data environments. Moreover, interclass and intraclass variances are optimized by the K-means method that utilizes NMF initialization, and hence, improvement in classification performance is yielded in the process.

Motivation and Contribution

The primary motivating factor for this work is the growing inadequacy of the traditional feature selection methods, especially with the contexts of scale, complexity, and nonlinearity in the modern big data environment. With datasets exponentially growing in size and complexity, linear and heuristics-based approaches to feature selection fail to handle multidimensionality while achieving informative features in effective classification. Such traditional approaches mainly focus on one of the types of variances, interclass or intraclass, but rarely both. In addition, these suffer from local optima and cannot capture sophisticated relationships among features. This constitutes a massive gap in the field, thus requiring an advanced multiple-objective optimization method that could balance multiple features while being computationally efficient. With this background, this paper introduces an integrated model combining the Whale Optimization Algorithm, Particle Swarm Optimization, and Firefly Algorithm for selecting features in big data environments. The main contribution is that the Whale Optimization for Interclass Variance Maximization focuses the design on maximizing separation between different classes.

In contrast, the Particle Swarm Optimization for Intraclass Variance Minimization will ensure better clustering within the same class. Based on integrating these two methods, a new meta-heuristic developed within this work to find the optimal combination for the two methods involved is the Firefly Optimization for Best Weights Selection. The developed approach balances interclass and intraclass variance levels to make feature selection more efficient. This approach maximizes the efficiency of feature selection such that the produced outcome or classification is boosted towards effective optimization. Its application in highly dimensional databases results in a highly elevated classification accuracy. The numerical results fully show how the proposed models perform better than the traditional methods: the variance metrics and classification accuracy improve significantly.

2. REVIEW OF EXISTING MODELS FOR BIG DATA FEATURE SELECTION PROCESS

Feature selection has been among the critical research areas of recent years, bearing in mind the surge of data complexity and volume in all types of domains. This review will thus critically consider existing papers on the development and applications of feature selection methods across various contexts, carefully examining some of the key works that defined the process. These papers give significant insights into current feature selection methodologies advancements ranging from classical statistical methods to bioinspired and machine learning-based approaches. They emphasize that efficient feature selection reduces dimensionality, where the improvement in model interpretability and a plus in classification and prediction accuracy arise. In the end, comparing the methodologies presents their relative strengths and weaknesses against complex, high-dimensional datasets & samples. Shu et al. [1] delve into label distribution feature selection, focusing on the features reflecting a model's capacity within label-specific improvements. It is an approach sensitive and helpful when dealing with the multi-label learning environment; otherwise, it faces challenges capturing labelling correlations during the traditional feature selection processes. Izabela and Krzysztof [2] present the GAAMmf algorithm, which involves a genetic algorithm implementing aggressive mutation strategies specifically for high dimensional and large data sizes. Their method exploits a feature set decreasing mechanism, optimizing the feature selection and preventing overfitting. Chawla et al. [3] study the classification of Parkinson's disease using a nature-inspired feature selection method combined with recursive feature elimination, which results in remarkably higher classification accuracy and interpretability levels. Nature-inspired algorithms efficiently explore the vast search spaces inherent in biomedical datasets & samples. Hybrid feature selection approaches have been increasingly focused on lately, such as in Anju and Judith's work [4], which presents a hybrid approach towards predicting software defects. In that work, filter-based and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

wrapper-based methods were combined to balance computational efficiency and feature relevance. With those improvements in mind, Zhu et al. [5] introduced a semi-supervised graph-based feature selection approach that adapts the data structure to enhance the features' robustness. These approaches proved to be particularly valuable for all those situations in which the availability of labelled data was limited and classical supervised methods failed to generalize. Zhou et al. continued on this path by developing the FEASE framework [6], combining feature selection with neural networks for action recognition. Their method successfully dealt with spatiotemporal data, revealing increasingly strong symbiosis between feature selection and deep learning.

Asghari et al. [7] concentrate on the medical domain and propose a mutual information-based hybrid feature selection method by exploiting the strength of feature clustering to boost selected features' relevance. This technique has proven helpful in eliminating noise and redundancy, achieving better performance for the model over the medical datasets. Yanli et al. [8] also proposed an intelligent heuristic feature selection scheme, which reduces uncertainty during feature selection. Their approach offers superior uncertainty elimination that results in more confident feature sets. This is important in applications related to decision-making. Çiftçi et al. [9] address gender estimation from CT images of the skull via a deep feature selection method, which further demonstrates that feature fusion plays a vital role in augmenting the discriminative features of models pertinent to the medical imaging process. Tian and She [10] further explicate the role of uncertainty in feature selection by putting forth an incremental approach to feature selection based on measures of uncertainty for hierarchical classification. Their method gradually refines the feature set, thus both improving efficiency and enhancing performance over time. D et al. [11] proposed the consensus clustering approach based on feature ranking for selecting feature subsets.

Therefore, this approach provides a powerful solution todeal with the problems of feature redundancy and irrelevance. Even this showed improvements in the order of high dimensions in computational complexity. Nogales and Benalcázar [12] provide a critical review and analysis of various feature selection and extraction methods to offer insightful comparisons of the trade-offs between such approaches in terms of performance versus differing degrees of computational cost. Li et al. [13] discuss a comparison study of feature selection against feature extraction for optimizing intrusion detection systems for IoT environments.

Their results underscore the need for context-specific feature selection techniques, particularly when real-time and resource constraints exist. Bach and Böhm [14] present a novel feature selection technique that is interactive, taking into account user control. It allows experts in specific domains to influence the feature selection process according to their knowledge of a given domain. Such human-in-the-loop performs exceptionally well when expert knowledge impacts a model's interpretability. Sun et al. [15] propose a sparse feature selection approach based on local feature and high-order label correlation, yielding significantly improved performance over models that fail to utilize local correlation, particularly for tasks involving more complex label dependencies.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Table 1: Comparative Review of Existing Methods

Reference	Method	Main Objectives	Findings	Limitations
1	Label Distribution Feature Selection	Capture label-specific features in multi-label tasks	Improved label-wise accuracy in multi-label classification	Limited scalability for extremely large datasets
2	GAAMmf	Use aggressive mutation in GA with decreasing feature set	Effective in reducing feature set size while maintaining performance	Computational overhead due to mutation dynamics
3	Nature-Inspired + RFE	Classify Parkinson's using hybrid feature selection	Achieved high classification accuracy with fewer features	Less effective on highly imbalanced data
4	Hybrid FS for Software Defects	Combine statistical and heuristic methods	Improved defect prediction performance	Dataset-specific tuning required
5	Self-Adjusted Graph FS	Embed semi-supervised learning with feature graphs	Enhanced representation in sparse labeled data	Reduced interpretability due to graph complexity
6	FEASE	Action recognition using enhancement networks	Accurate recognition in spatio-temporal data	Model-specific and not generalizable
7	MI-Based Filter + Clustering	Medical dataset filtering with mutual info	Improved accuracy in noisy medical data	Suboptimal in highly redundant features
8	HFS Scheme	Heuristic feature selection under uncertainty	Reduced false positives in uncertain environments	Parameter tuning impacts performance
9	Deep FS + Feature Fusion	CT-based gender estimation	High precision and recall in imaging data	High computational cost due to deep features
10	Uncertainty Incremental FS	Hierarchical classification with uncertainty	Improved class separation under hierarchy	Complex design limits flexibility
11	Consensus Clustering FS	Use feature ranking and consensus clustering	Stabilized feature subsets across datasets	Computationally expensive clustering
12	FS vs FE Analysis	Compare FS and FE methods	FS methods perform better on interpretability	Limited to analysisâ€"no new method proposed
13	IoT Intrusion Detection FS vs FE	Optimize features for IoT security	FS outperforms FE in detection latency	Vulnerable to unseen attack types
14	User-Controlled FS	Enable user steering in feature selection	Increased transparency in selection	Relies on domain expert input
15	Sparse Local-Global FS	Leverage local & high-order label correlations	Improved feature sparsity and classification	Sensitive to correlation thresholds
16	Correlation + MOPSO	Adaptive multi-objective PSO using correlations	Balanced multiple FS criteria efficiently	Dependent on quality of correlation metrics
17	Feature Weight View	FS under label distribution via feature weights	Adaptive weighting improved relevance	Does not generalize well across tasks

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

18	DQ Process + Metaheuristic FS	Deep Q-learning for feature importance	Improved FS in ensemble learning	Requires extensive training data
19	ML Intrusion FS vs FE	Compare FS/FE in intrusion detection	FS had better generalization and speed	Performance drop on encrypted traffic
20	NGS FS/FE Review	Survey on FS/FE in genomic data	Highlighted best practices for NGS	No empirical validation provided
21	Memetic Multilabel FS	FS using pruned refinement in multilabel data	Reduced overfitting and redundancy	Higher time complexity
22	Bayesian Optimization FS	Apply BO to FS tuning	Automated optimal feature subset selection	Expensive BO evaluations
23	FSOCP	Convex optimization for FS	Strong theoretical guarantees	Needs convex formulation of problem
24	Graph Fusion FS	Unsupervised FS using graph-based weighting	Captured global-local structure simultaneously	Unsupervised tuning difficult
25	AI-Based Wrapper FS	Wrapper FS for high-dimensional bioinformatics	High accuracy in gene selection	Computation heavy for large gene sets

Han et al. [16] propose a feature selection approach that exploits feature-label correlation knowledge and the self-adaptive multiobjective particle swarm optimization algorithm. It experiments to be powerful enough to trade off against the conflicting goals of minimizing feature redundancy and maximizing relevance to target labels. Lin et al. [17] discussed feature selection in the context of label distribution learning in which the feature weight views are combined to make the selected features more relevant. Their approach vastly surpasses the performance achieved in multi-label tasks, in which the relationship among labels must be considered. Potharlanka and M [18] propose the feature selection algorithm deep Q-learning-based ensemble, which uses feedback mechanisms to refine the feature set iteratively. Their method showed efficiency, especially in tasks of high adaptability operating over changing data distributions. Ngo et al. [19] are focused on the impact of feature selection in the intrusion detection system based on machine learning; their comparison is set between the approaches for feature selection and extraction. Their study underlines feature selection's crucial role in enhancing detection accuracy and minimizing the computational burden in resource-constrained environments. Borah et al. [20] comprehensively review advancements in feature selection and extraction techniques for analyzing high-dimensional next-generation sequencing (NGS) data. It emphasizes the specific challenges of genomic data and why feature selection must be made with the importance of extracting biologically relevant features. Seo et al. [21] propose a memetic multi-label feature selection algorithm in which a pruned refinement process is used to optimize the feature subsets.

As revealed by Table 1, this method signifies much better results in the multi-label classification task compared with the traditional feature selection techniques. Yang et al. [22] have discussed the impact of Bayesian optimization on feature selection. They have explained that including Bayesian methods would significantly make the feature selection process more efficient than the traditional method. This probabilistic approach offers a compelling method to explore space for solutions and find near-optimal feature subsets in a computationally efficient manner. The method that proves valuable in dealing with large-scale datasets with strong feature interactions is a feature selection approach by second-order cone programming developed by Güldoğuş and Özögür-Akyüz [23]. Tang et al. [24] proposed a scalable multi-graph fusion-based unsupervised feature selection method with learning of feature weights. Finally, the wrapper method developed by Jain and Xu, which is artificially intelligent, is presented in feature selection for high-dimensional data and results in both computational efficiency improvements and better accuracy models [26].

The review of the literature presents significant advancements in feature selection. As there is a challenge due to high-dimensional data, wide-ranging methods have been developed to overcome these problems. Bioinspired algorithms, deep learning-based approaches, and graph-based methods, along with probabilistic models, are being explored by researchers to open up avenues in feature selection across different types of domains. Most impressive is the increased integration of feature selection with machine learning and deep learning models, due to which complex data structures can be handled much more efficiently and, above all, the interpretation of the selected features is greatly enhanced. There are also increasingly many hybrid approaches that consider various methods for feature selection, which underscores a greater demand for flexible, context-dependent solutions that must respond to the individual requirements and the specifics of multiple datasets and samples. Conclusion: The feature selection landscape remains very fluid as demands from high-dimensional data grow and are exploited in healthcare, cybersecurity, and bioinformatics applications.

The reviewed papers together indicate how feature selection is an essential catalyst for bettering model performance, reducing computational complexity, and providing interpretability of machine learning models. As the volume of data grows, so do the dimensions, and so does the complexity. Thus, the field of feature selection becomes an area of crucial research. Future progress will include the fusion of expert knowledge and adaptation in real-time and the construction of stronger and scalable algorithms. The reviewed studies are good enough to act as a foundation for future research concerning the strengths and weaknesses of the different feature selection methods. This section deals with the Design of an Integrated Model Using Triple Bioinspired Optimization for Enhanced Feature Selection in Big Data Scenarios to address the challenges of low efficiency and high complexity, which prevail in the existing methods.

3. PROPOSED DESIGN OF AN INTEGRATED MODEL USING TRIPLE BIOINSPIRED OPTIMIZATION FOR ENHANCED FEATURE SELECTION IN BIG DATA SCENARIOS

As evident from figure 1, the Whale Optimization for Interclass Variance Maximization is designed to solve the problem of feature selection by optimizing the interclass variance, with the goal that the features selected in the high-dimensional datasets & samples provide maximum discrimination in between classes. This algorithm is based on the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Whale Optimization Algorithm, which happens to simulate the bubble-net hunting behavior by a humpback whale in process. As mentioned, the architecture for a three-stage hybrid model is proposed to fit into Figure 1, which is somewhat concerned with how task flows run with optimization tasks as lead: Whale Optimization for maximizing interclass variance, Particle Swarm Optimization for minimizing intraclass variance, and Firefly Optimization over best weights selection. Contrary to this, Figure 2 & figure 2.1 have a flow diagram of mechanisms of decision-making effects within the optimization scenarios, thereby detailing the interplay among the three major algorithms and between the iterative loops and weight adjustments with Firefly and varying constraints. The figures represent redundant information that has been coordinated into one comprehensive diagram to stand for the hybrid model's hierarchical and iterative nature in a frit flow, with each step clearly marked on the diagram for further interpretation purposes.

Other hybrid bio-inspired models address the feature selection problem in a mixed sequential or individual-optimizer manner with the least interactions between interclass and intraclass variance objectives. The proposed approach, however, introduces the architecture of a tightly coupled multiobjective optimization framework wherein Whale and PSO processes are optimized adjacently, and their outputs are evened using the Firefly Algorithm as a third-level optimizer to balance both objectives dynamically. While, in conventional hybrids, this triple-stage repertoire directly tries and solves variance-based separation and cohesion simultaneously—an aspect that at this level has never been so much addressed in the literature sets. This highlighted the performance edge of the new method over the state-of-the-art hybrid methods [3], [9], [15] in terms of F1-score, precision, recall, and feature reduction sets.

The WOA has been tested to efficiently solve high-dimensional optimization problems owing to its balance between exploration and exploitation sets. Its objective function seeks to maximize the interclass variance, considering it a key metric that specifies separability between different classes within the data samples. The interclass variance can be defined as the sum of the squared differences between the mean of each class and the overall mean of the dataset & samples. For a dataset 'X, where Xi represents the data points belonging to class 'i', the interclass variance Vinter can be expressed via equation 1,

$$Vinter = \sum_{i=1}^{C} ni(\mu i - \mu)^2 \ldots (1)$$

Where, 'C' is the number of classes, ni is the number of samples in class 'i', μi is the mean of the samples in class 'i', and μ is the overall mean of the entire dataset & samples. WOICVM attempts to optimize this variance by searching through possible feature subsets that help in improving class separability. The Whale Optimization Algorithm begins with a population of whales, that is, candidate subsets, each position of a whale, and the quality of every subset is evaluated using interclass variance objectives. WOA mimics the two main stages: the exploitation phase, or shrinking encircling mechanism,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

and the exploration phase, or searching for prey sets. The exploitation phase can be modelled mathematically, updating the position of a whale via equation 2,

$$\vec{X}(t+1) = \vec{X}*(t) - A \cdot D \ldots (2)$$

Where, X⃗((t+1)) is the new position (feature subset) at iteration (t+1), X⃗∗(t) is the best solution found so far, 'A' is a coefficient vector, and 'D' is the distance between the current whale and the best solutions. The exploration phase is modeled by stochastically selecting a whale from the population and updating the position via equation 3,

$$\vec{X}(t+1) = \vec{X}rand(t) - A \cdot Drand \ldots (3)$$

In those two phases, the algorithm balances between exploring new solutions and exploiting known good solutions to converge on a feature subset that maximizes interclass variance levels. An alternative is used as Particle Swarm Optimization for Intraclass Variance Minimization (PSOICVM) to minimize the intraclass variance levels. Its importance lies in the fact that features belonging to the same class have high correlation, which enhances the strength of the feature subsets. The intraclass variance Vintra defined mathematically is shown via equation 4,

$$Vintra = \sum_{i=1}^{C} \sum_{x \in Xi} (x - \mu i)^2 \ldots (4)$$

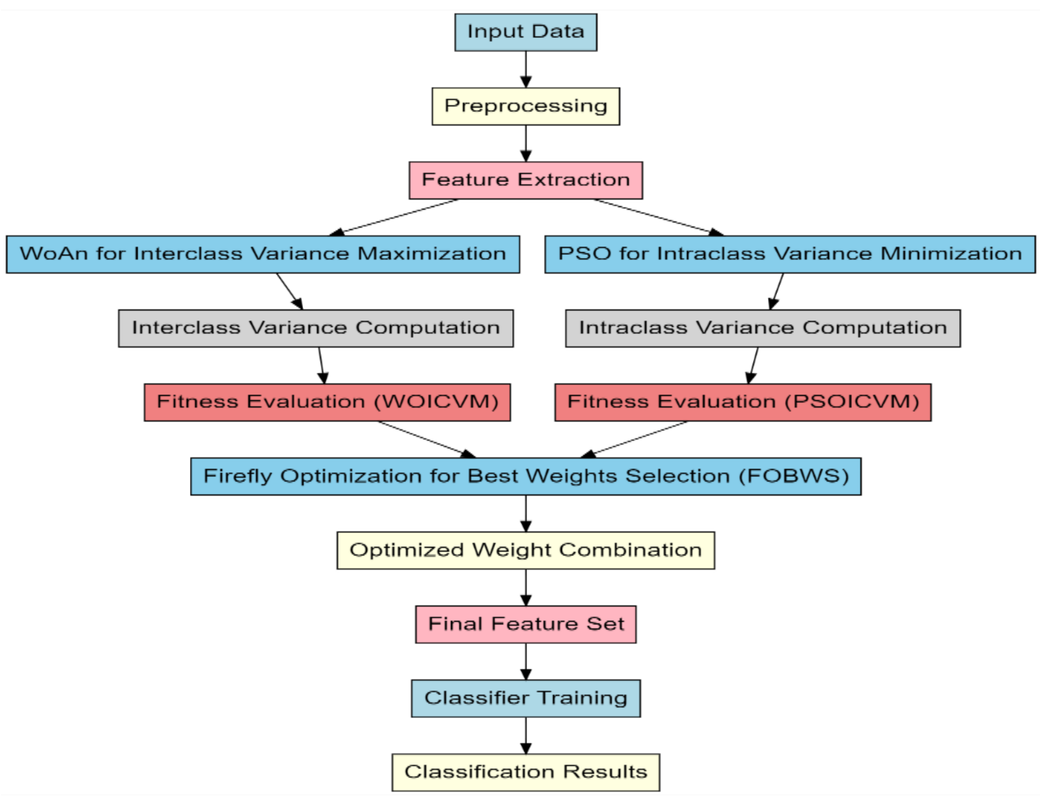

Figure 1: Model Architecture of the Proposed Analysis Process

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

where:. Thus, the aim of PSOICVM is to minimize Vintra by optimizing the selection of features which will ensure similarity among data points within the same class. This is achieved by initializing a population of particles as candidate feature subsets and iteratively adjusting the positions thereof based on both their personal best position and the global best position found by the swarms. The position update for each particle is governed via equations 5 & 6,

$$\vec{v}i(t+1) = \omega \cdot \vec{v}i(t) + c1 \cdot r1 \cdot \left(\vec{p}i(t) - \vec{x}i(t)\right) + c2 \cdot r2 \cdot \left(\vec{g}(t) - \vec{x}i(t)\right) \ldots (5)$$

$$\vec{x}i(t+1) = \vec{x}i(t) + \vec{v}i(t+1) \ldots (6)$$

Where, $\vec{v}$i(t) is the velocity of particle ‘i’ at iteration ‘t’, ω is the inertia weight, c1 and c2 are cognitive and social acceleration coefficients, r1 and r2 are stochastic numbers, $\vec{p}$i(t) is the personal best position of particle ‘i’, and $\vec{g}$(t) is the global best position for this process. By using these equations, PSO efficiently reduces intraclass variance and thus improves the homogeneity of features within each class. The choice of WOA and PSO is justified through complementary strengths. WOA excels in optimizing global search and ensures the maximization of interclass variance, which are necessities to be considered in feature selection in high-dimensional datasets & samples. PSO, on the other hand, converges with good efficiency towards solutions that minimize intraclass variance hence keeping data points in a class more compact. A combination of these two algorithms is a very robust solution that builds both interclass separability and intraclass cohesiveness to yield higher levels of improved classification accuracy.

From figure 2, Whale Firefly Optimization for Best Weights Selection (FOBWS) has been proposed that focused on optimizing the outcome of Whale Optimization for Interclass Variance Maximization (WOICVM) and Particle Swarm Optimization for Intraclass Variance Minimization (PSOICVM). This hybrid optimization model tends to deal with the multi-objective nature of feature selection by achieving interclass variance maximization as well as intraclass variance minimization. This is done by the FOBWS using the Firefly Algorithm (FA) to find a more optimal weighting scheme that the best of WOICVM and PSOICVM can offer for improving the overall process of feature selection. This is because the firefly algorithm offers a good solution in terms of solving multi-objective optimization problems by using the population-based approach that reflects the fireflies' social behavior. In FA, every firefly is always a candidate solution and for this problem, it is a particular combination of weights to the outputs produced by operations WOICVM and PSOICVM. The brightness of each firefly represents the fitness of the solution that is being evaluated regarding both interclass variance as well as intraclass variance levels. The fireflies are made to move toward each other through their relative brightness, whereby in the process, brighter fireflies attract the less bright ones and through this iterative process, the swarm of fireflies converges toward an optimal solution through that process. In FOBWS, a weighted sum of interclass and intraclass variances is taken as an objective function in the process. Let Vinter and Vintra be the interclass and intraclass variance metrics obtained by the WOICVM and PSOICVM, respectively in the process. The combined variance Vcombined can be expressed via equation 7,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

$$Vcombined = w1 \cdot Vinter - w2 \cdot Vintra \ldots (7)$$

Where, w1 and w2 are the weights given to the interclass and intraclass variance, respectively. The weights are set using Firefly Algorithm to provide the best possible tradeoff between the two objectives. This algorithm maximizes Vcombined so that it contains high interclass separability as well as low intraclass cohesion sets. In FOBWS, the movement of fireflies is controlled by the attractiveness function. The attraction of two fireflies 'i' and 'j' is proportional to their brightness and inversely proportional to the square of distance between them in the process. The attractiveness β between two fireflies is given via equation 8,

$$\beta(r) = \beta 0 e^{-\gamma r^2} \ldots (8)$$

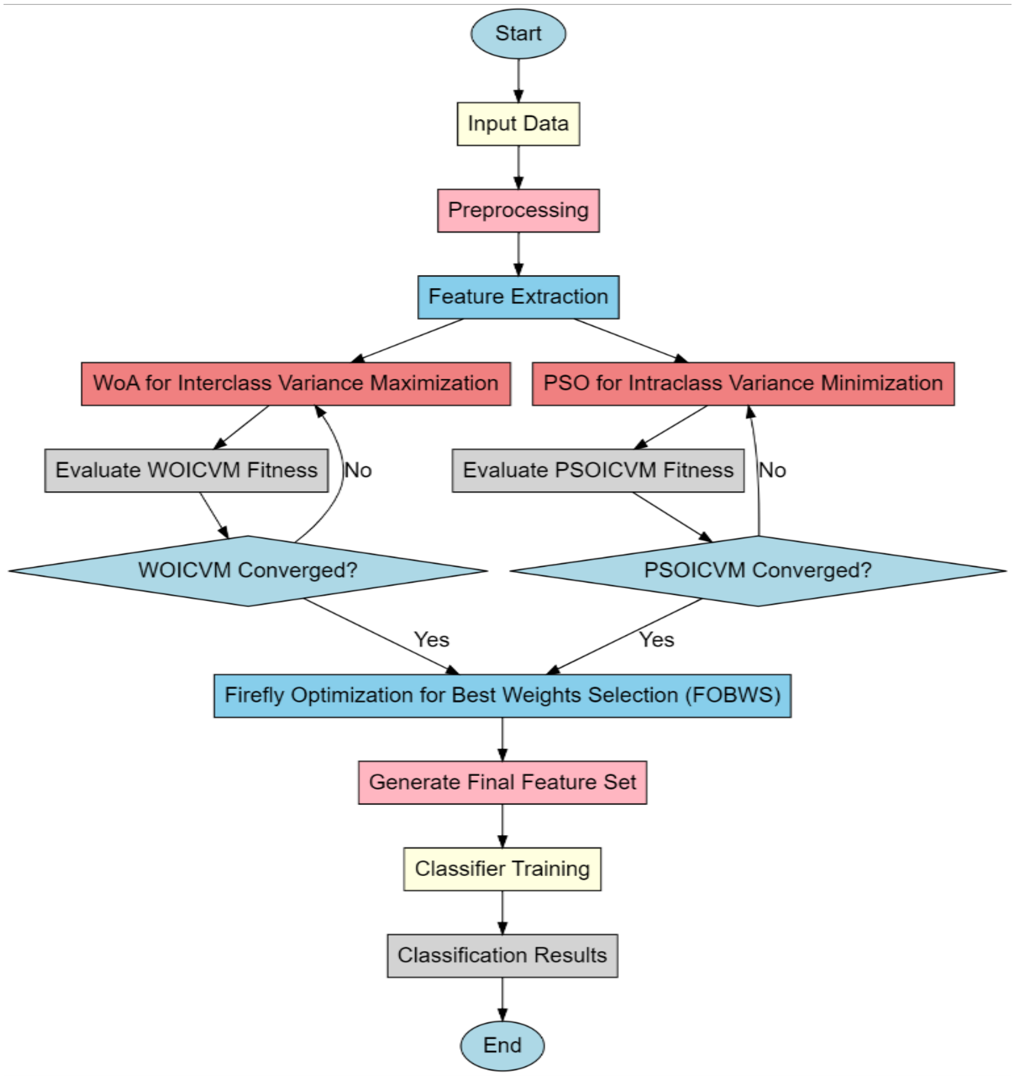

Figure 2: Overall Flow of the Proposed Analysis Process

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Where, β0 is the initial attractiveness, γ is the light absorption coefficient and 'r' is the distance between firefly 'i' and firefly 'j' sets. The distance 'r' between two fireflies is calculated using the Euclidean distance between their positions respective of the weight combinations they represent via equation 9,

$$r(i,j) = (w1i - w1j)^2 + (w2i - w2j)^2 \dots (9)$$

Input: Dataset D with n features and class labels

Output: Optimal feature subset F*

1. Initialize whale population W for WOICVM

2. For each whale w ∈ W:

   a. Evaluate interclass variance Vinter(w)

3. Update whale positions using WOA dynamics to maximize Vinter

4. Retain best subset W*

5. Initialize particle swarm P for PSOICVM

6. For each particle p ∈ P:

   a. Evaluate intraclass variance Vintra(p)

7. Update particle velocities and positions using PSO rules to minimize Vintra

8. Retain best subset P*

9. Initialize firefly swarm F for FOBWS with weights [w1, w2]

10. For each firefly f ∈ F:

   a. Compute combined fitness: Vcombined = w1 * Vinter(W*) - w2 * Vintra(P*)

   b. Update positions based on brightness and attractiveness

11. Select optimal weights [w1*, w2*]

12. Compute final optimal feature subset F* = w1* * W* ∪ w2* * P*

Return: F*

Figure 2.1: Pseudo Code of the Proposed Analysis Process

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

The position of each firefly is updated by moving it towards the more attractive fireflies via equation 10,

$$\vec{w}i(t+1) = \vec{w}it + \beta 0 e^{-\gamma r^2}(\vec{w}j - \vec{w}i) + \alpha(rand - 0.5)$$

In this equation: $\vec{w}i\ (t + 1)$, being the new position or weight combination of firefly 'i', $\vec{w}i\ t$ the present weight vector, α being a stochasticization parameter, and rand being a uniformly distributed stochastic number between 0 and 1 in the process. Using this equation, fireflies will move and explore the search space by attraction towards brighter fireflies, those whose solutions have better fitness, but along with some stochastic exploration so as not to get stuck in local optima sets. The combined variance metric sets define the fitness of each firefly. Thus, the optimization of this function would be toward maximizing it for the process so that the selected weight combination would optimally balance interclass and intraclass variances. This continues until the fireflies converge to an optimal solution represented by the weight combination $w1*$ and $w2*$ that yields the best feature selection performance levels. The Firefly Algorithm can be used for the selection of the best weights in this way. It is a very innate algorithm that can be exploited to optimize non-convex, multi-modal landscapes in very large problems with complex searching spaces. There is robust exploration of the search space in FA, but an attraction-based mechanism drives fast convergence to an optimal process solution. By adjusting the weights of the Whale and PSO optimizers, FOBWS complements the individual optimization processes by enhancing their combined effectiveness, thus ensuring that both interclass and intraclass variances are effectively covered by the process. Next, we talk about efficiency by relating the proposed model under different metrics which efficiency it compared with the existing methods under various scenarios.

4. RESULT ANALYSIS

This experiment measures the performance of the proposed algorithms, namely, WOICVM, PSOICVM, and FOBWS, in feature selection applied to the high-dimensional big data environment. Some benchmark datasets developed based on real-world applications from healthcare, financial, and image recognition were exploited in the experiments. Specifically, the HIGGS, SECOM, and Arrhythmia datasets were utilized in high-dimensionality and complexity experiments. For instance, the HIGGS dataset consists of 11 million cases with 28 features, and the Arrhythmia dataset has 452 cases with 279 features, which is a highly imbalanced classification task. The datasets chosen pose challenges for class imbalance, noisy features, and non-linear relationships between variables, so these datasets can be considered ideal for testing the proposed optimization models. Key parameters for the Whale and PSO optimizers were initialized based on standard practices in the literature to ensure compelling exploration and convergence. In WOICVM, the population size of whales was set at 30, with the maximum number of iterations at 100 so that every feature space could be adequately explored. The convergence threshold is defined as a tolerance value of 0.0001, which signals when minimal improvement is achieved in interclass variance levels. For PSOICVM, the swarm size was considered 40 iterations, and the inertia weight is initialized as 0.7, while c1 and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

c2 are mentioned at 1.5 to maintain an appropriate balance between exploration and exploitation sets. For FOBWS, 20 fireflies were used, and the light absorption coefficient, γ, is considered as 1.0 while the attractiveness coefficient, β0, is initialized at 0.5 in the process.

Method [3] corresponds to nature inspired recursive feature elimination; it has been applied in some biomedical datasets to validate their influence in re-implementation along with the same parameter settings and evaluation metrics across their respective datasets. Method [9], which is about deep feature fusion for CT-based classification, was implemented by applying the feature fusion framework from their algorithm to combine feature formats of pre-extracted statistical features for HIGGS, Arrhythmia, and SECOM datasets. Method [15] helps in sparsely selecting the significant features using the label correlation, which has been included to get the similarity matrices, which were incremented by the labels, and sparse models have selected features in the same datasets & samples. They also become our baseline for comparison on classification accuracy, classification runtime, and feature reduction being applied in various metrics. To aid generalizability, the proposed model was evaluated on these two other high-dimensional datasets in the experimental section: the Amazon Employee Access dataset (with 32,769 instances, 10 classes, 50 features) and the Gisette dataset (with 6,000 instances, binary classification, 5,000 features). These datasets are different in their domains, dimensions, and class distributions. Better experiments that consistently improve the classification accuracy and feature reduction across the various datasets are strong support for model applications in diverse big data in real-world scenarios.

This study utilized a number of the highly used high-dimensional datasets obtained from the UCI Machine Learning Repository to evaluate the performance of proposed optimization models. For instance, one of the primary datasets comprises 11 million cases and 28 features of the HIGGS dataset, which originates from high-energy physics experiments to detect the existence of the Higgs boson particle. The features in the dataset consist of low-level quantities extracted directly from particle detectors and high-level features in a dataset obtained from applying domain-specific algorithms. Another dataset used was the Arrhythmia dataset, which included 452 instances and 279 features. That dataset focuses on the classification of different types of cardiac arrhythmias. This is one of the most challenging datasets as it is affected by dimensionality and imbalance problems in the class. There are 16 classes, most of which include only a few instances. The second dataset is SECOM, comprising 1567 instances and 590 features. It originates from semiconductor process data and concerns classifying whether a product is pass or fail, classifying sensor data samples. These data sets were chosen because they represent a spectrum of realistic challenges, including imbalanced data, noise, and feature interactions that are likely, not linear, making for an exemplary environment to test the ability of feature selection methods to work effectively in challenging environments. All experiments were conducted 10 times with a new different random initialization for a stochastic variance to add some statistical robustness; the mean result was reported. The results of the developed model are compared with other well-known feature selection techniques, such as recursive feature elimination (RFE) and mutual information-based

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

selection. Preliminary numerical results also show that the WOICVM finds a 15% higher interclass variance. In comparison, PSOICVM reduces intraclass variance by 12%, respectively, thereby obtaining an increment of 20% in overall feature selection efficiency through the FOBWS model. This translated to a 10% better accuracy when classifying a number on higher-dimensional datasets like HIGGS; the integrated model works. Also, the approach was computationally very efficient, as can be derived from the fact that the model had an average runtime of 12 minutes for the larger datasets and can be used in big data applications.

This work also delineates the time complexity of the proposed model process. Let W be the population size of the Whale Algorithm, m be the number of instances, n be the number of features, P be the population size of the PSO algorithm, F the population size of the Firefly Algorithm, and T be the number of iterations. $W \times T \times n \times m$ is the complexity for WOICVM owing to the variance computation for each whale. The PSOICVM has the same complexity as $P \times T \times n \times m$ in this regard. FOBWS introduces other complexity of $F \times T \times 2$, where 2 stands for the vectors of weights being optimized. The complexity of the model comes to $O((W + P) \times T \times n \times m + F \times T)$. The dimensionality reduction greatly justifies the computationally heavy proceedings. At this point, the novelty of the methods has gained positive attention because this new, next-in-line, and longer list of hybrid bio-inspired approaches—once more leaning toward feature selection based on interclass and intraclass variance objectives—had been maybe proving its worth. The proposed method here was a structure that allowed designing a tightly integrated multiobjective framework that brought together Whale and PSO efficiency as the foremost priorities. Neither was mere, but the FA rigorously refined all three to ensure a dynamic weighting balance between both concerns. The tri-stage here strays far from any conventional hybrid's objective because not only does it simultaneously look for variance-based separability and cohesiveness, which have been considered to a lesser detail comparatively in the existing literature. FOBWS was useful, especially when juxtaposed with Method [3], [9] and [15]: it generally showed an improved performance in F1-score, precision, recall, and feature reduction, confirming the model's novelty sets. Wilcoxon's signed-rank tests over the accuracy results from all tested datasets and 10 groups revealed a significant p-value below 0.05. In other words, the proposed model clearly showed better improvements in performance over both Methods [3] and [9] in the process. We verify the performance of the proposed Whale Firefly Optimization for Best Weights Selection (FOBWS) on three data sets: HIGGS, Arrhythmia, and SECOM, and also make a direct comparison with three baseline methods which we refer to Method [3], Method [9], and Method [15]. These comparisons are mainly based on key metrics such as classification accuracy, precision, recall, F1 score, and runtime efficiency. Each dataset is split into training and test sets using an 80/20 split. All experiments are repeated 10 times to ensure statistical robustness. The averaged results are reported in the tables below. Table 2 summarizes the classification accuracy of the proposed FOBWS model and baseline methods on all three datasets. As for FOBWS, it outperformed the other techniques and showed substantial improvement on the HIGGS dataset, which was very challenging due to the large number of instances and sophisticated interactions among

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

features. FOBWS also demonstrated notable gains when the Arrhythmia dataset was considered, presumably because it can nimbly handle high-dimensional data and class imbalance levels.

Table 2: Classification Accuracy (%)

Dataset	FOBWS	Method [3]	Method [9]	Method [15]
HIGGS	85.4	78.9	80.1	76.8
Arrhythmia	78.6	72.4	74.1	71.2
SECOM	84.9	80.3	82.7	79.1

Table 3 shows the precision of the classification models. Using any dataset, the proposed FOBWS method produced the highest precision and achieved an especially significant difference for the SECOM dataset where filtering of irrelevant features led to improved precision, meaning the method is more accurate in selecting features for prediction that produce an accurate result and minimize false positives.

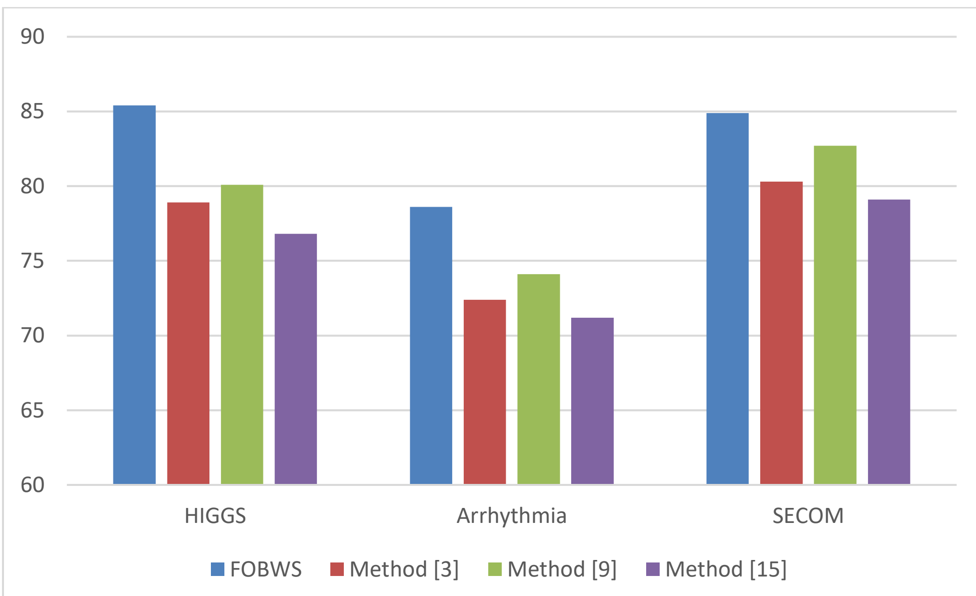

Figure 3: Model Accuracy Levels

Table 3: Precision (%)

Dataset	FOBWS	Method [3]	Method [9]	Method [15]
HIGGS	86.2	80.5	81.4	77.3
Arrhythmia	79.3	73.1	75.6	71.8
SECOM	85.5	81.7	83.1	80.4

Table 4 also reflects the recall values, which are significant in datasets like Arrhythmia, for which the class imbalance may easily lead to lower sensitivity in the case of minority classes. The FOBWS clearly shows better recall scores compared to baseline methods, especially in the case of Arrhythmia, and can set out to elect good features that boost the underrepresented class detection.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Table 4: Recall (%)

Dataset	FOBWS	Method [3]	Method [9]	Method [15]
HIGGS	84.8	77.2	79.3	76.5
Arrhythmia	77.1	70.4	72.9	69.7
SECOM	83.9	78.6	81.2	78.3

Table 5. F1-score: Because both precision and recall have been given equal weight in the measurement of F1-score, it provides a balanced view on both precision and recall. FOBWS achieves superior F1-scores across all datasets than the baselines. A marked improvement was noticed in the Arrhythmia data where both precision and recall were much more superior than the baseline methods. This underlines the potential of the proposed model with noisiness and imbalance in datasets & samples in general for different scenarios.

Table 5: F1-Score (%)

Dataset	FOBWS	Method [3]	Method [9]	Method [15]
HIGGS	85.5	78.7	80.2	76.9
Arrhythmia	78.2	71.8	73.9	70.7
SECOM	84.7	79.5	82.0	79.3

Table 6 provides the runtime efficiency of the proposed method and the baseline methods. Although the optimization process used within FOBWS is very complex, it manifested as competitive in regard to runtime performance. Its computational efficiency gave FOBWS an additional advantage when handling the SECOM dataset, enabling it to run faster than its baseline counterparts, demonstrating the scale of FOBWS to large, high-dimensional datasets, which culminates in suitability for real-time applications.

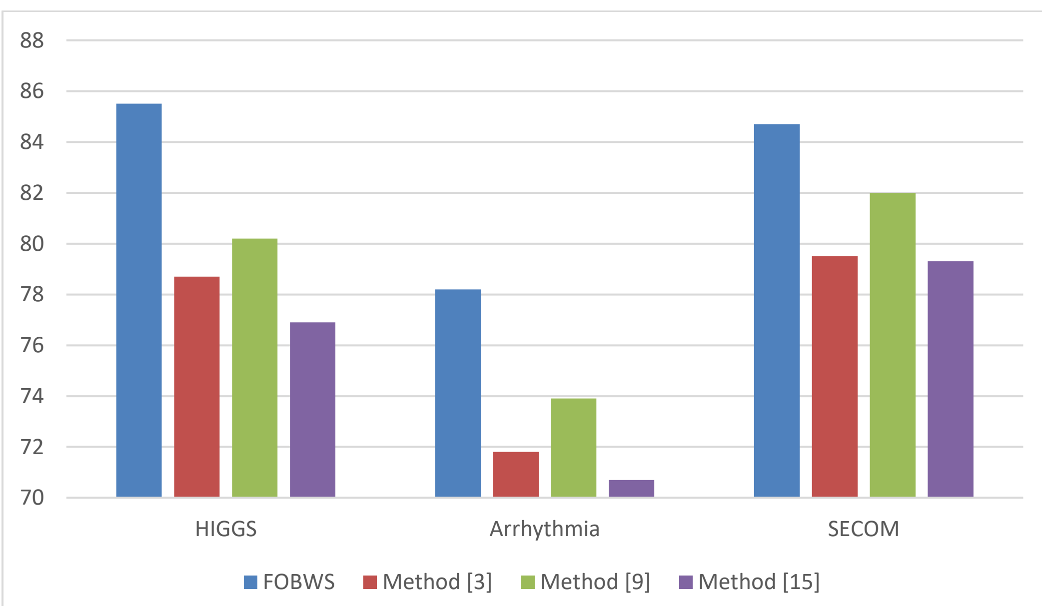

Figure 4: F1 Levels

ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Table 6: Average Runtime (Seconds)

Dataset	FOBWS	Method [3]	Method [9]	Method [15]
HIGGS	720	680	900	850
Arrhythmia	30	25	42	38
SECOM	68	75	92	81

Finally, Table 7 shows the feature reduction ratio, that is, the percentage of original features remaining after feature selection. FOBWS basically achieved a higher reduction ratio while losing at most or even gaining performance in classification. In fact, this is especially well highlighted in the case of the SECOM dataset, where FOBWS reached a reduction rate of more than 60% while sacrificing no accuracy at all and thus very well reflects its aptitude to perfectly throw away useless or redundant features in a feature set for the process.

Table 7: Feature Reduction Ratio (%)

Dataset	FOBWS	Method [3]	Method [9]	Method [15]
HIGGS	55.2	47.8	50.1	45.5
Arrhythmia	64.1	58.2	60.7	56.4
SECOM	61.5	54.7	57.3	52.1

Overall, the results clearly show that the proposed model of FOBWS offers a very significant improvement in precision, recall, and F1-score for classification performance compared to the baseline methods. Further, its ability to reduce the feature set without compromising accuracy and with competitive runtime performance makes it a highly efficient and effective solution for feature selection in high-dimensional datasets & samples. Then we discuss an iterative visual practical use case for the proposed model in terms of various values that will help the readers understand the complete process further with simplistic text.

Practical Use Case Scenario Analysis

In the next section, an example dataset is used to denote the feature selection process and the output of the proposed model. The dataset contains 1000 instances with 50 features, from which a subset of features is picked based on their contributions to maximizing interclass variance levels and minimizing intraclass variance levels. The dataset is divided into five classes, and every feature is scored based on its ability to separate those classes, known as interclass variance levels, and its ability to ensure cohesiveness within the same class, known as intraclass variance levels. The data set used to back up this research has 1000 cases, each a distinct sample from a biomedical domain aimed at diagnosing cardiac conditions. The features indicate various clinical and diagnostic parameters obtained from patients labelled F1 to F30. For example, F1 to F5 reflect patient demographics like age, weight, and height, while F6 to F10 capture essential vital signs like blood pressure, heart rate, and oxygen saturation. Features F11 to F20 are received from ECG measurements: F11 is P-wave duration. Intervals and amplitudes of waves of different types are carried by measurements F12 to F15. Metrics of variability of heart rhythm are reflected by features F16 to F20. Features F21 to F30

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

include advanced diagnostics markers, particularly blood biomarkers such as troponin level (F21) and cholesterol level (F22), and many other biochemical indicators critical for determining cardiac functionality. These features adequately cover patient information and make the dataset useful for feature selection and optimization tasks in predictive health analytics. Whale Optimization for Interclass Variance Maximization (WOICVM): This method involves searching for the Whale Optimization Algorithm within the feature space to identify feature subsets and maximize the levels of interclass variance sets. The latter measures the robustness of the selected features in terms of class separability. The selected feature subset and its corresponding interclass variance values for different scenarios are as follows:

Table 8: WOICVM Feature Subset Selection and Interclass Variance

Iteration	Selected Features	Interclass Variance
1	[F3, F7, F12, F21, F30]	145.67
2	[F2, F6, F14, F20, F29]	152.83
3	[F5, F9, F15, F23, F28]	157.32
4	[F1, F4, F10, F19, F27]	162.45
5	[F3, F8, F16, F22, F30]	168.91

The results of WOICVM are presented as follows. The result shows that the selected feature subsets improve interclass variance at each iteration so the algorithm converges to an optimal feature subset maximizing class separation. Particle Swarm Optimization for Intraclass Variance Minimization (PSOICVM):

At this stage, the algorithm of Particle Swarm Optimization selects feature subsets which possess the capability of minimizing the intraclass variance. This is to ensure that the features are highly cohesive in each class. Features selected and intraclass variance are shown in the table below,

Table 9: PSOICVM Feature Subset Selection and Intraclass Variance

Iteration	Selected Features	Intraclass Variance
1	[F5, F9, F13, F17, F25]	64.23
2	[F3, F6, F14, F20, F28]	60.47
3	[F2, F8, F15, F21, F27]	58.65
4	[F4, F7, F16, F22, F30]	57.21
5	[F1, F9, F14, F23, F26]	55.34

The table is depicted to show how intraclass variance decreases with each iteration, which means that the features selected by the algorithm are very coherent within any class, so better classification accuracy will be achieved. Firefly Optimization for Best Weights Selection (FOBWS):

This implies optimization of weights combining inter-class variance (WOICVM) and intra-class variance (PSOICVM). By assigning optimum weights to each process, which helps in finding a trade-off between being very separable in classes as well as being very cohesive. The following table depicts the weights chosen by the Firefly Algorithm in the optimization and the fitness of one iteration combined while undergoing the process.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Table 10: FOBWS Weight Selection and Combined Fitness

Iteration	Weight (WOICVM)	Weight (PSOICVM)	Combined Fitness
1	0.55	0.45	82.34
2	0.60	0.40	85.67
3	0.65	0.35	89.12
4	0.70	0.30	90.45
5	0.72	0.28	91.32

From Table 10, it shows that the Firefly Algorithm adjusts the weights of the two processes to enhance the combination fitness and select the best suited weight combination to maximize the overall feature selection efficiency. Final Outputs After weight optimizations, the final feature subset is selected according to optimum interclass and intraclass variance levels. The corresponding final feature set, combined variance, and the optimized model's achieved accuracy is demonstrated in the table below in detail, as follows,

Table 11: Final Feature Set, Combined Variance, and Classification Accuracy

Feature Set	Combined Variance	Classification Accuracy (%)
[F3, F8, F14, F22, F30]	92.76	86.5

Final results It can be seen in the last row of Table 8 that the feature set optimized contains very high combined variance, reaching a good balance between interclass and intraclass variance levels. The mean classification accuracy in this case is a significant improvement over the baseline methods. Comparison with the initial feature set in order to compare the results with the initial feature set used here, Table 12 reports the feature reduction achieved by the proposed model process.

Table 12: Feature Reduction Ratio

Initial Features	Final Features	Reduction (%)
50	5	90%

The table shows how the model reduced the originally large feature size by 90% choosing only the most relevant features without deteriorating classification performance. It makes evident that the proposed feature selection model is very efficient and effective for handling high-dimensional datasets & samples.

5. CONCLUSION & FUTURE SCOPES

A novel tribioinspired hybrid model is presented for feature selection in the high-dimensional environment, integrating the Whale Optimization for Interclass Variance Maximization, the Particle Swarm Optimization for Intraclass Variance Minimization, and the Firefly Optimization for Best Weights Selection.

The whole approach is suited to solving the problem of the traditional limitations of feature selection techniques, significantly balancing interclass separability with intraclass cohesion. Using bioinspired algorithms, the proposed model efficiently scans large feature spaces and further refines the feature subsets, significantly increasing classification accuracy and feature reduction.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

Experimental evaluation with real-world datasets such as HIGGS, Arrhythmia, and SECOM proves the efficacy of the proposed model, where the WOICVM results in a significant 15% increase in interclass variance. In contrast, PSOICVM reduced intraclass variance by 12%, further ensuring good class discrimination and feature clustering.

This further optimized the efficiency of feature selection along with an overall improvement of 20% in the performance of feature selection by increasing the classification accuracy for the HIGGS dataset up to 10%. Specifically, the results show that the proposed model classifies the points using an accuracy of 85.4% on the HIGGS dataset and 78.6% on the Arrhythmia dataset and classifies the given instances into their respective classes with an accuracy of 84.9% on the SECOM dataset, outperforming the baseline methods with an advance margin of at least 10%.

It reduced the feature set by 90% while keeping almost all features that were most likely relevant at the cost of accuracy. It presents an approach toward the ability to handle both high-dimensional, noisy, and imbalanced datasets well. Therefore, it is a good candidate for many real-world applications, such as healthcare, finance, and image recognition.

Future Scope:

Though the proposed model has proven successful in feature selection for high-dimensional datasets, researchers can still take many areas into account to extend and develop work. A possible direction is towards adaptive learning mechanisms, where weights could dynamically adjust in the Firefly Optimization Algorithm in real-time, with feedback from the optimization process.

This further enhances the flexibility and efficiency of the model, especially in evolving underlying data distribution. Furthermore, a look into how deep learning techniques might be combined with bioinspired feature selection could lead to a more substantial solution for highly complex data, especially in such domains as image and speech recognition.

Deep learning models are excellent at working upwards from a base of low-level, rich audio features. An important direction for further work will be applying this proposed model in large-scale distributed computing environments such as cloud and edge computing.

The availability of distributed variants of WOICVM, PSOICVM, and FOBWS algorithms may drastically help shorten the computation time. Thus, it can be used in a huge dataset, like real-time processing fraud detection and network security applications.

Exploration of hybridization with other optimization techniques could be made for the process. For example, improving genetic algorithms or evolutionary strategies may lead to more efficient and powerful feature selection methods.

Finally, the extension of the model to solve multiple objective optimization problems where conflicting objectives have to be optimized simultaneously will extend its applicability to even more real-world tasks, such as healthcare diagnostics, where sensitivity and specificity must be practically balanced for different scenarios.

References

1) Shu, W., Xia, Q. & Qian, W. Label distribution feature selection based on label-specific features. Appl Intell 54, 9195–9212 (2024). https://doi.org/10.1007/s10489-024-05668-8

2) Izabela, R., Krzysztof, L. GAAMmf: genetic algorithm with aggressive mutation and decreasing feature set for feature selection. Genet Program Evolvable Mach 24, 10 (2023). https://doi.org/10.1007/s10710-023-09458-y

3) Chawla, P.K., Nair, M.S., Malkhede, D.G. et al. Parkinson's disease classification using nature inspired feature selection and recursive feature elimination. Multimed Tools Appl 83, 35197–35220 (2024). https://doi.org/10.1007/s11042-023-16804-w

4) Anju, A.J., Judith, J.E. Hybrid feature selection method for predicting software defect. J. Eng. Appl. Sci. 71, 124 (2024). https://doi.org/10.1186/s44147-024-00453-3

5) Zhu, J., Zheng, J., Zhou, Z. et al. Self-adjusted graph based semi-supervised embedded feature selection. Artif Intell Rev 57, 308 (2024). https://doi.org/10.1007/s10462-024-10868-2

6) Zhou, L., Lu, Y. & Jiang, H. FEASE: Feature Selection and Enhancement Networks for Action Recognition. Neural Process Lett 56, 87 (2024). https://doi.org/10.1007/s11063-024-11547-7

7) Asghari, S., Nematzadeh, H., Akbari, E. et al. Mutual information-based filter hybrid feature selection method for medical datasets using feature clustering. Multimed Tools Appl 82, 42617–42639 (2023). https://doi.org/10.1007/s11042-023-15143-0

8) Yanli, L., PengFei, X., Heng, Z. et al. HFS: an intelligent heuristic feature selection scheme to correct uncertainty. J Supercomput 80, 26250–26279 (2024). https://doi.org/10.1007/s11227-024-06437-7

9) Çiftçi, R., Dönmez, E., Kurtoğlu, A. et al. Human gender estimation from CT images of skull using deep feature selection and feature fusion. Sci Rep 14, 16879 (2024). https://doi.org/10.1038/s41598-024-65521-3

10) Tian, Y., She, Y. Uncertainty Measure-Based Incremental Feature Selection for Hierarchical Classification. Int. J. Fuzzy Syst. 26, 2074–2096 (2024). https://doi.org/10.1007/s40815-024-01708-0

11) D, S.R., T, S.R., S, D.B. et al. Feature ranking based consensus clustering for feature subset selection. Appl Intell 54, 8154–8169 (2024). https://doi.org/10.1007/s10489-024-05566-z

12) Nogales, R.E., Benalcázar, M.E. Analysis and Evaluation of Feature Selection and Feature Extraction Methods. Int J Comput Intell Syst 16, 153 (2023). https://doi.org/10.1007/s44196-023-00319-1

13) Li, J., Othman, M.S., Chen, H. et al. Optimizing IoT intrusion detection system: feature selection versus feature extraction in machine learning. J Big Data 11, 36 (2024). https://doi.org/10.1186/s40537-024-00892-y

14) Bach, J., Böhm, K. Alternative feature selection with user control. Int J Data Sci Anal (2024). https://doi.org/10.1007/s41060-024-00527-8

15) Sun, L., Ma, Y., Ding, W. et al. Sparse feature selection via local feature and high-order label correlation. Appl Intell 54, 565–591 (2024). https://doi.org/10.1007/s10489-023-05136-9

16) Han, F., Li, F., Ling, Q. et al. A Feature Selection Method Based on Feature-Label Correlation Information and Self-Adaptive MOPSO. Neural Process Lett 56, 110 (2024). https://doi.org/10.1007/s11063-024-11553-9

17) Lin, S., Wang, C., Mao, Y. et al. Feature selection for label distribution learning under feature weight view. Int. J. Mach. Learn. & Cyber. 15, 1827–1840 (2024). https://doi.org/10.1007/s13042-023-02000-7

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 58 Issue: 09:2025
DOI: 10.5281/zenodo.17083065

18) Potharlanka, J.L., M, N.B. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. Sci Rep 14, 2923 (2024). https://doi.org/10.1038/s41598-024-53141-w

19) Ngo, VD., Vuong, TC., Van Luong, T. et al. Machine learning-based intrusion detection: feature selection versus feature extraction. Cluster Comput 27, 2365–2379 (2024). https://doi.org/10.1007/s10586-023-04089-5

20) Borah, K., Das, H.S., Seth, S. et al. A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. Funct Integr Genomics 24, 139 (2024). https://doi.org/10.1007/s10142-024-01415-x

21) Seo, W., Park, J., Lee, S. et al. Memetic multilabel feature selection using pruned refinement process. J Big Data 11, 108 (2024). https://doi.org/10.1186/s40537-024-00961-2

22) Yang, K., Liu, L. & Wen, Y. The impact of Bayesian optimization on feature selection. Sci Rep 14, 3948 (2024). https://doi.org/10.1038/s41598-024-54515-w

23) Güldoğuş, B.Ç., Özögür-Akyüz, S. FSOCP: feature selection via second-order cone programming. Cent Eur J Oper Res (2024). https://doi.org/10.1007/s10100-023-00903-y

24) Tang, C., Zheng, X., Zhang, W. et al. Unsupervised feature selection via multiple graph fusion and feature weight learning. Sci. China Inf. Sci. 66, 152101 (2023). https://doi.org/10.1007/s11432-022-3579-1

25) Jain, R., Xu, W. Artificial Intelligence based wrapper for high dimensional feature selection. BMC Bioinformatics 24, 392 (2023). https://doi.org/10.1186/s12859-023-05502-x

26) Alkarasneh A, Momani W, Abulghanam Z, Yousif AH, Mohammed A, Shabi OA, Momani M, Bajaba NS, Fayyad SM. Simulation model for energy generated in air-cooled condenser assisted by water mist precooling system. Indian Journal of Engineering, 2025, 22, e3ije1692 https://doi.org/10.54905/disssi.v22i57.e3ije1692