A Bio-Inspired Earthworm Optimization Algorithm Combined with PCA for Improved Feature Selection in Machine Learning Models

Amit Kumar Saxena
Professor
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

Damodar Patel
Research Scholar
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

Umesh Kumar Shriwas
Research Scholar
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

Abhishek Dubey
Assistant Professor
College of Computing and Information Sciences, Information Technology Department
University of Technology and Applied Sciences, Salalah, Sultanate of Oman

Gayatri Sahu
Research Scholar
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

Shreya Chinde
Research Scholar
Department of CSIT
Guru Ghasidas Vishwavidyalaya
Bilaspur, India

ABSTRACT
High-dimensional data often leads to increased computational complexity and reduced model performance due to the curse of dimensionality. This study introduces an effective feature selection and classification framework that integrates the Earthworm Optimization Algorithm (EWA), Principal Component Analysis (PCA), and supervised classifiers, K Nearest Neighbors (KNN) and Support Vector Machine (SVM). EWA, a bio-inspired metaheuristic based on the foraging behavior of earthworms, efficiently identifies optimal feature subsets. PCA is then applied to further minimize dimensionality while preserving essential variance. The proposed EWA-PCA was evaluated on 19 benchmark datasets using stratified 10-fold cross-validation and standard classification metrics. In the KNN average accuracy of 19 datasets, using the original feature set achieved 77.65% of accuracy, while the EWA-PCA achieved better 86.56%; similarly, in SVM, 84.43% of accuracy was achieved in the original feature, while the EWA-PCA achieved 88.10%. Results show that EWA-PCA consistently outperforms conventional and modern feature selection techniques, including Chi2, ReliefF, SIFS, mRMR, ATFS, and EmPo. EWA-PCA achieved better classification accuracies with KNN and SVM, demonstrating high stability and substantial feature reduction. The findings validate EWA-PCA as a scalable, accurate, and efficient solution for high-dimensional data classification.

Keywords
Feature selection, Earthworm Optimization Algorithm, Principal Component Analysis, Dimensionality reduction.

1. INTRODUCTION
The vast amount of input features in high-dimensional data [1] has made machine learning [2] difficult owing to its rapid increase. To increase algorithm efficiency and forecast accuracy, feature selection (FS) [3,4] is a critical pre-processing step that eliminates redundant, noisy, or unnecessary input. It enhances model simplicity, generalization, and learning speed. FS techniques involve trade-offs between subset size and performance, satisfying evaluation necessities, and enhancing an evaluation measure. While Subset Evaluation selects feature subsets using a search approach, Individual Evaluation ranks features according to their significance. Selecting appropriate features is essential to prevent issues in models, particularly when managing noisy or irrelevant data and when the feature count surpasses the sample size.

Dimensionality reduction [5] uses feature extraction [6] or FS to eliminate duplicate and noisy features. Through the use of methods like Principal Component Analysis (PCA) [7], Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA), feature extraction combines original features and converts data into a lower-dimensional space. Distinguish between FS techniques that are independent of classifiers and those that are dependent on classifiers. Filter methods [8] assess the significance of features by analyzing their inherent properties, like statistical correlations with the output variable, independently of any learning algorithm, like Laplacian score [9], variance, chi-squared, Cosine similarity [10], and mutual information [11], etc. Wrapper methods [12] utilize a predictive model to assess feature subsets by training and testing the model across various combinations. While they can be computationally intensive, they have the potential to identify optimized subsets effectively. Embedded methods [13] integrate FS directly into the model training process. Examples include decision trees or regularization techniques like Lasso. They achieve a balance between computational efficiency and predictive power, although their applicability may be limited in

some contexts [14].

This paper presents a supervised model that integrates machine learning, dimensionality reduction, and optimization for effective FS and classification. The Earthworm Optimization Algorithm (EWA) [15] is employed to identify the most relevant features that enhance model performance. Principal Component Analysis (PCA) is applied to further reduce dimensionality while preserving essential variance. Classification is conducted using Support Vector Machine (SVM) and K Nearest Neighbors (KNN) with stratified k-fold cross-validation. Performance is evaluated through accuracy, confusion matrices, and classification reports. Comparative analysis using bar charts and PCA projections demonstrates improved classification accuracy (CA) and reduced computational complexity for high-dimensional data. The contributions of this work are as follows:

- Introduce a hybrid feature selection method that combines Earthworm Optimization Algorithm (EWA) and Principal Component Analysis (PCA) to optimally select the most relevant features and reduce dimensionality while preserving classification performance
- The proposed EWA-PCA method employs cross-validation with KNN/SVM classifiers to evaluate each candidate feature subset, ensuring that only high-quality feature sets are retained, ultimately boosting predictive accuracy.
- By evaluating performance on a broad mix of 19 datasets, some binary and others multi-class, that demonstrate the method's stability and adaptability. Robustness is evidenced through consistently high accuracy metrics across varied data types and complexities.
- We benchmarked the proposed EWA–PCA against eight well-known feature-selection techniques. Across 19 datasets, the proposed framework consistently outperformed all competitors in classification accuracy, showcasing its superior efficacy.

The paper is structured in the following manner. Section 2 provides a concise overview of various ensemble FS techniques. Section 3 presents the EWA-PCA. Section 4 illustrates the effectiveness and efficiency of the method through comprehensive experiments, in Section 5 describes the Results and Discussion. The last section contains the Conclusion and Future Work.

2. Literature Review

The objective of FS is to remove a significant number of irrelevant and redundant features. Methods for FS that utilize filtering techniques demonstrate remarkable efficiency and are adept at managing high-dimensional datasets with speed. In supervised learning, these techniques evaluate features according to their significance in relation to class labels. Common techniques for ranking features encompass the Pearson correlation coefficient and present a framework based on consensus groups aimed at enhancing the stability of FS in high-dimensional, small-sample datasets.

ReliefF [16]. To obtain stable and reliable FS in high-dimensional data, this research suggests an ensemble FS method that evaluates the dependability of individual feature pickers.

The mRMR [17] has established itself as a fundamental method in the domain of FS. By skillfully integrating the principles of maximum relevance and minimum redundancy through mutual information, it addresses critical challenges related to the analysis of high-dimensional datasets.

The SIFS [18] technique offers a robust method for supervised feature ranking in high-dimensional datasets. Through the integration of label information within a graph-based framework and the application of infinite path analysis to evaluate feature significance, SIFS enhances the detection of informative and non-redundant features for classification tasks.

ATFS [19] is specially designed for complex high-dimensional settings. Their method improves the selection process by integrating rapid, non-dominated sorting with ensemble learning techniques, resulting in greater robustness and flexibility.

The proposed approach, EmPo [20], presents a compelling and resilient strategy for feature reduction. This method effectively tackles several frequently opposing goals at once while utilizing a variety of different FS techniques, offering a thorough resolution to the challenges posed by high-dimensional data. As a result, it constitutes an important advancement in the evolution of optimization-focused FS techniques within the realm of machine learning.

Wu et al. employed five established filter FS algorithms, including Chi-square [21] and the F test, for FS.

The FSM [22] method improves classification by integrating various feature subsets to obtain complementary information. This fusion-based method enhances robustness, minimizes redundancy, and guarantees improved generalization, representing a significant advancement in ensemble FS for high-dimensional data classification challenges.

3. Proposed Method

In this section present the structure of the proposed FS and classification approach based on the Earthworm Optimization Algorithm (EWA). To enhance prediction performance, the technique combines FS, classification, class imbalance management, and data preparation. The EWA-PCA for FS and performance measurement is shown in Figure 1. The input dataset is first gathered and normalized to make sure all features are on the same scale, which is necessary for accurate classifier performance. To identify the most relevant features after normalization, the Earthworm Optimization Algorithm (EWA), a metaheuristic algorithm inspired by natural phenomena, is employed. SVM and KNN are two classifiers used to guide this optimization process and assess the efficacy of specific feature subsets. The objective is to select the best features in order to enhance CA. Principal Component Analysis (PCA) is used to further reduce dimensionality by combining related features into a smaller collection of uncorrelated components after EWA has completed the initial selection. This two-step reduction procedure improves the model's performance and computing efficiency. Finally, use the refined set of features to evaluate classifier performance using standard metrics, thereby validating the effectiveness of the proposed approach.

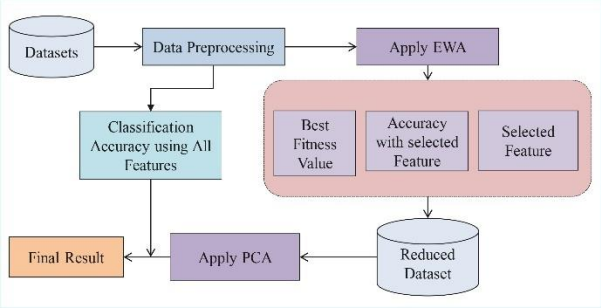

Fig 1: Flow chart of EWA-PCA

3.1 Earthworm Optimization Algorithm (EWA)

The EWA is a metaheuristic optimization method inspired by nature that imitates how earthworms travel through soil. It is intended to solve intricate optimization issues by mimicking the coordinated movements, expansions, contractions, and food-seeking behaviors of earthworms. This algorithm's exploration and exploitation capabilities make it especially useful for solving high-dimensional and nonlinear optimization problems. In EWA, earthworms are shown as possible solutions in a search space. Mathematical formulae that balance exploration (finding new answers) and exploitation (building upon current best solutions) govern each earthworm's position and movement. The random walk and the guided movement are the two movements that the algorithm mainly simulates. While controlled movement helps earthworms to converge towards favorable locations based on prior results, the random walk gives them more freedom to explore the search space, thus increasing the optimization process's efficiency.

Expansion-Contraction: To investigate new areas, earthworms travel at random within a predetermined range:

$$X_i^{t+1} = X_i^t + \alpha \cdot (\text{rand} - 0.5) \quad (1)$$

where Xit+1 is the earthworm's updated location, Xit is its current position, α is the step size, and $rand$ is a random value between 0 and 1.

Peristaltic Movement: Earthworms move about according to the most effective solutions they discover:

$$X_i^{t+1} = X_i^t + \beta \cdot (X^t_{best} - X_i^t) \quad (2)$$

where X^t_{best} is the best solution found so far, and β\betaβ is a contraction coefficient that controls convergence. By dynamically modifying search agents, avoiding local optima, and providing global convergence, EWA efficiently resolves optimization issues. It is appropriate for FS, machine learning, and engineering applications because of its simplicity and reliability. EWA performs better in high-dimensional optimization challenges by using mathematical modeling and population-based search methodologies. All things considered, it offers a successful, naturally inspired method for accurately and computationally efficiently resolving challenging optimization problems.

Experimented with several EWA settings, including step sizes of 0.1, 0.01, and 0.001 and population sizes ranging from 10 to 100 (in increments of 10) in order to improve the parameter selections. For the majority of datasets, the best results were consistently obtained with a population size of 60 and a step size of 0.1.

Table1. EWA Parameter

Description	Value
Population size	60
Maximum iteration	50
Step size	0.1
Patience	10
Early Stopping	10
Classifier	SVM (Kernel = 'rbf') and KNN (k = 3)

Adaptive step size: Decreases over iterations using:

Adaptive step size: Decreases over iterations using:

$$\mathbf{adaptive_{step}} = \mathbf{step_{size}} \times \left(1 - \frac{\mathbf{iteration}}{\mathbf{max_{iter}}}\right) \quad (3)$$

Probability transformation with a sigmoid function:

$$\boldsymbol{P} = \frac{1}{1+e^{new\_solution}} \quad (4)$$

Termination Condition: Either max_iter is reached or early stopping is triggered.

These parameters control the exploration-exploitation balance in the Earthworm Optimization process for FS.

3.2 Principal Component Analysis (PCA)

A popular method for resolving these problems is Principal Component Analysis (PCA), which converts the information into a lower-dimensional space while maintaining the majority of the variation. To maximize model performance and minimize feature dimensionality while maintaining crucial information, the provided method applies PCA to a chosen subset of features. Only the chosen features are designated as *X*selected. *X*selected are extracted from the dataset at the start of the operation. The PCA transformation uses these features as its input, selecting them based on their significance in the classification task. PCA is used with a variance retention criterion of 95% to ensure substantial data representation. This means that the bare minimum of Principal components needed to account for 95% of the total variance is automatically chosen. The transformed feature set *X*PCA is produced mathematically as follows:

$$X_{PCA} = X_{Selected}.W \quad (5)$$

where W represents the eigenvectors corresponding to the largest eigenvalues of the covariance matrix of *X*selected, the number of principal components retained, ncomponent, is dynamically determined by PCA. The classifier reduces the number of features while maintaining good accuracy by retaining 95% variance. The PCA-based transformation is a useful method for FS and classification tasks since it also increases interpretability and computational efficiency.

4. EXPERIMENTS

4.1 Experiment design

Experiments were conducted utilizing 19 commonly employed benchmark datasets. A greater value signifies that a larger number of features have been effectively eliminated. Prior to executing the experiments, it is essential to establish the optimal value for comparison in the tests. Jupyter version 7.2.2

is used to conduct the tests on a desktop computer running Windows 10 Pro and equipped with an Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz, 3292 MHz, 2 Cores, 2 Logical Processors, and 8.00 GB of installed physical memory (RAM).

4.2 Machine Learning Classifiers

K-Nearest Neighbors (KNN): KNN [23] is a non-parametric, instance-based learning algorithm used for both classification and regression. It classifies a data point by identifying the majority class among its k nearest neighbors, determined using distance metrics such as Euclidean, Manhattan, or Minkowski. As a lazy learner, KNN does not train an explicit model but instead retains the entire dataset and performs computations during prediction. The choice of k significantly impacts the model's performance by controlling the trade-off between sensitivity to noise and the smoothness of the decision boundary [24].

Support Vector Machines (SVM): SVM) are robust supervised learning algorithms applied to both classification and regression problems. They operate by constructing an optimal hyperplane that maximizes the margin between different classes in the training data. SVMs can handle non-linearly separable data through the use of kernel functions, such as polynomial and radial basis function (RBF) kernels. These kernels project the input data into higher-dimensional spaces, enabling the creation of flexible and non-linear decision boundaries, making SVMs particularly effective for complex classification tasks[25].

4.3 Experimental datasets

The 19 datasets utilized from the machine learning repository (https://archive.ics.uci.edu/datasets) in the studies are briefly described in Table 2. In these datasets, there are between 34 and 10000 features, 60 and 7797 samples, and 2 to 26 categories.

Table 2. Dataset Details

Datasets	Features	Instances	Categories
ALLAML	7129	72	2
Arcene	10000	200	2
BASEHOCK	4862	1993	2
COIL20	1024	1440	20
COLON	2000	62	2
GLI_85	22283	85	2
Ionosphere	34	351	2
ISOLET	617	7797	26
Lung	3312	203	5
Nci9	9712	60	9
ORL	1024	400	40
Orlraws10P	10304	100	10
PCMAC	3289	1943	2
Prostate_GE	5966	102	2
RELATHE	4322	1427	2
sonar	60	208	2
WarpAR10P	2400	130	10
WarpPIE10P	2420	210	10
Yale	1024	165	15

5. RESULTS AND DISCUSSION

Table 3 presents a comparison of CA between the original feature set and the selected feature set of EWA and proposed EWA-PCA using KNN and SVM classifiers. For the KNN classifier, an average CA of 77.65% was achieved using all features, and in the selected feature of the EWA method, a CA of 85.16%, while the selected features of the proposed EWA-PCA method yielded a significantly higher average CA of 86.56%. Similarly, with the SVM classifier, the average CA using all features was 84.43%, and selected features using EWA was 87.46% while selected features of the proposed EWA-PCA method improved to 88.10%.

In the comparison of original features and selected features by EWA-PCA in KNN, all datasets showed improved CA when using the selected features. For SVM, sixteen out of the total datasets demonstrated better CA with the selected features. However, in the ALLAML dataset, the CA slightly decreased from 92.86% with all features to 90.8% using the selected features. The CA of the selected features from EWA alone is lower than that of the proposed EWA-PCA method. This highlights the importance of applying PCA, indicating that the EWA-PCA approach achieves better performance compared to EWA alone.

The improvements are particularly noticeable at the dataset level in complex and high-dimensional datasets. For instance, using both EWA and EWA-PCA increased the KNN accuracy in the Arcene dataset from 56% (original) to 89.50%. EWA-PCA scored 73.33% in the Nci9 dataset, which is much better than 46.67% using original features. In a same similar manner, EWA-PCA improved KNN accuracy in ALLAML from 77.86% (original) to 86.11%. The proposed method outperformed the initial 98.10% accuracy for SVM on WarpPIE10P, achieving 100% accuracy.

As shown in Colon, PCMAC, RELATHE, and several others, EWA-PCA often beat EWA alone, even if EWA alone frequently enhanced performance. This shows how the Earthworm Optimization Algorithm and PCA work together to effectively remove redundant data while maintaining discriminative information. The standard EWA performed slightly better than EWA-PCA in a few datasets (such as GLI_85 and Prostate_GE), indicating that PCA may sometimes eliminate weakly relevant variance. However, in most datasets, the proposed EWA-PCA approach consistently produced the best or nearly best results. Overall, the results indicate that the EWA-PCA algorithm effectively selects important features and removes redundant ones, leading to enhanced classification performance in most cases.

The comparison of the performance metrics of the original features and the proposed EWA-PCA approach using both KNN and SVM classifiers across various datasets is shown in Tables 4 and 5. The proposed method produces significant improvements in accuracy, precision, recall, and F1-score for both classifiers compared to the original feature set. For instance, in the KNN classifier on the ALLAML dataset, the F1-score improved from 82.22% to 88.89%, recall improved from 78.72% to 85.11%, and precision from 86.05% to 93.02% using EWA-PCA. Across all datasets, similar patterns are seen, with EWA-PCA continuously preserving or improving the balance between recall and precision, which eventually results in higher F1-scores.

Table 3: Comparison of the CA between the original datasets and the Selected Features using KNN and SVM classifiers

Datasets	CA of Original Features using KNN	CA of Selected Features using KNN with EWA	CA of Selected Features using KNN with EWA-PCA	CA of Original Features using SVM	CA of Selected Features using SVM with EWA	CA of Selected Features using SVM with EWA-PCA
ALLAML	77.86%	79.17%	**86.11%**	**92.86%**	87.50%	90.28%
Arcene	56%	89.50%	**89.50%**	56%	84.00%	**83.00%**
BASEHOCK	71.30%	**83.29%**	82.29%	95.74%	**95.89%**	95.53%
COIL20	99.17%	100%	**99.79%**	94.79%	**100%**	99.93%
COLON	78.57%	85.48%	**87.10%**	83.81%	88.71%	**90.32%**
GLI_85	85.83%	84.71%	**90.59%**	**90.56%**	85.88%	88.24%
Ionosphere	84.90%	**91.43%**	91.14%	91.17%	94.30%	95.16%
ISOLET	82.88%	87.18%	**87.24%**	95%	**96.60%**	96.35%
Lung	95.60%	98.03%	**98.03%**	87.69%	95.07%	**95.07%**
Nci9	46.67%	68.33%	**73.33%**	8.33%	**30.00%**	25%
ORL	89.75%	92.50%	**92.75%**	95.50%	96.75%	**96.75%**
Orlraws10P	93%	94%	**96%**	99%	99%	**99%**
PCMAC	70.72%	81.68%	**82.30%**	87.49%	87.50%	**90.28%**
Prostate_GE	83.18%	87.25%	**89.22%**	93.09%	91.18%	**93.14%**
RELATHE	79.33%	82.83%	**82.83%**	87.69%	90.28%	**91.10%**
sonar	80.79%	88.94%	**90.38%**	82.69%	85.10%	**87.02%**
WarpAR10P	46.92%	57.69%	**60%**	**90.19%**	77.69%	80.77%
WarpPIE10P	94.29%	**98.71%**	98.10%	98.10%	100%	**100%**
Yale	58.60%	67.27%	**67.88%**	74.38%	76.36%	**76.97%**
Average	77.65%	85.16%	**86.56%**	84.43%	87.46%	**88.10%**

Table 4: Comparison of the Performance metrics between the original datasets and the Selected Features using KNN classifiers

Datasets	Performance metrics of Original Features				Performance metrics of Selected Features with EWA-PCA			
	CA	Precision	Recall	F1-Score	CA	Precision	Recall	F1-Score
ALLAML	77.86%	86.05%	78.72%	82.22%	86.11%	93.02%	85.11%	88.89%
Arcene	56%	61.76%	56.25%	58.89%	89.50%	91.74%	89.29%	90.49%
BASEHOCK	71.30%	71.41%	71.27%	71.34%	82.29%	82.36%	82.28%	82.32%
COIL20	99.17%	99.58%	99.60%	99.56%	99.79%	99.72%	98.91%	99.65%
COLON	78.57%	86.49%	80.00%	83.12%	87.10%	92.11%	87.50%	89.74%
GLI_85	85.83%	92.73%	86.44%	89.47%	90.59%	96.36%	89.83%	92.98%
Ionosphere	84.90%	90.95%	84.89%	87.82%	91.14%	94.91%	91.11%	92.97%
ISOLET	82.88%	87.38%	85.96%	85.94%	87.24%	86.52%	87.55%	87.01%
Lung	95.60%	95.76%	94.87%	95.96%	98.03%	98.25%	97.03%	98.93%
Nci9	46.67%	47.00%	46.44%	48.50%	73.33%	75.14%	73.40%	74.20%
ORL	89.75%	89.80%	88.99%	88.73%	92.75%	93.40%	92.10%	94.00%
Orlraws10P	93%	93%	93%	93%	96%	96%	96%	96%
PCMAC	70.72%	80.80%	68.30%	74.10%	82.30%	95.10%	79.00%	86.30%
Prostate_GE	83.18%	84.31%	82.69%	83.50%	89.22%	90.20%	88.46%	89.32%
RELATHE	79.33%	82.18%	79.33%	80.73%	82.83%	85.32%	82.80%	84.04%
sonar	80.79%	82.57%	81.08%	81.82%	90.38%	91.74%	90.09%	90.91%
WarpAR10P	46.92%	46.80%	46.94%	46.70%	60%	61.12%	60.15%	59.42%
WarpPIE10P	94.29%	93.29%	94.48%	93.81%	98.10%	98.50%	98.17%	98.91%
YALE	58.60%	58.67%	56.76%	57.67%	67.88%	67.48%	69.08%	66.47%
Average	77.65%	80.55%	77.68%	79.10%	86.56%	88.89%	86.20%	87.50%

Table 5: Comparison of the Performance metrics between the original datasets and the Selected Features using SVM classifiers

Datasets	Performance metrics of Original Features				Performance metrics of Selected Features with EWA-PCA			
	CA	Precision	Recall	F1-Score	CA	Precision	Recall	F1-Score
ALLAML	92.86%	95.65%	93.62%	94.62%	90.28%	95.45%	89.36%	92.31%
Arcene	56%	61.76%	56.25%	58.89%	83.00%	86.11%	83.04%	84.55%
BASEHOCK	95.74%	95.79%	95.70%	95.74%	96.14%	96.19%	96.10%	96.14%
COIL20	94.79%	94.20%	94.39%	95.00%	99.93%	99.90%	98.98%	99.82%
COLON	83.81%	89.47%	85.00%	87.18%	90.32%	94.74%	90.00%	92.31%
GLI_85	90.56%	96.36%	89.83%	92.98%	88.24%	94.55%	88.14%	91.23%
Ionosphere	91.17%	94.91%	91.11%	92.97%	95.16%	97.27%	95.11%	96.18%
ISOLET	95%	95.80%	95.20%	94.29%	96.35%	94.80%	96.15%	96.00%
Lung	87.69%	84.49%	87.19%	85.86%	95.07%	95.15%	95.76%	96.03%
Nci9	8.33%	8.02%	7.56%	8.01%	25%	24.62%	25.50%	24.52%
ORL	95.50%	94.50%	95.16%	94..87%	96.75%	96.55%	95.89%	96.71%
Orlraws10P	99%	99.10%	99.24%	98.54%	99%	99.21%	99.07%	99.23%
PCMAC	87.49%	96.90%	88.10%	92.30%	90.28%	99.10%	91.60%	95.20%
Prostate_GE	93.09%	94.12%	92.31%	93.20%	93.14%	94.12%	92.31%	93.20%
RELATHE	87.69%	89.52%	87.68%	88.59%	91.10%	92.45%	91.14%	91.79%
sonar	82.69%	84.40%	82.88%	83.64%	87.02%	88.18%	87.39%	87.78%
WarpAR10P	90.19%	90.69%	98.52%	89..83%	80.77%	80.15%	82.77%	81.38%
WarpPIE10P	98.10%	98.19%	99.10%	98.57%	100%	99.57%	99.91%	99.81%
YALE	74.38%	74.48%	75.08%	74..23%	76.97%	76.91%	75.55%	75.97%
Average	84.43	86.23	84.94	71.60	88.13	89.74	88.09	88.96

Table 6: Comparison of CA (In %) Between Proposed EWA-PCA with Eight Other FS Method using KNN classifier

Datasets	PCA	ReliefF	mRMR	Chi2	SIFS	ATFS	EmPo	FSM	EWA-PCA
ALLAML	70.89	88.75	**98.57**	97.14	76.43	95.89	90.18	82.14	86.11
Arcene	83.5	88	78	85	85.5	56	69.5	86	**89.5**
BASEHOCK	76.92	73.11	93.88	91.92	66.63	88.81	84.4	**89.21**	82.29
COIL20	99.58	99.65	99.65	98.96	99.65	-	97.36	99.65	**99.79**
COLON	73.57	**90.48**	86.9	86.9	78.81	86.9	88.81	85.24	87.1
GLI_85	84.58	90.56	95.28	92.92	83.61	**91.94**	81.53	89.31	90.59
Ionosphere	85.21	87.18	85.48	87.47	83.18	85.47	84.63	90.33	**91.14**
ISOLET	82.76	72.44	77.5	83.33	67.63	-	29.04	85.58	**87.12**
Lung	95.12	96.07	95.71	95.6	92.14	-	89.74	96.57	**98.03**
Nci9	30	50	56.67	56.67	50	-	38.33	53.33	**73.33**
ORL	89.75	92.75	91.25	90.25	88.75	-	76.75	91	**92.75**
Orlraws10P	92	96	95	**99**	90	-	91	94	96
PCMAC	69.38	64.08	**89.01**	84.61	61.97	75.91	70.2	82.24	82.3
Prostate_GE	84.18	**93.09**	92.09	92.03	76.45	91.09	91.09	89.18	89.22
RELATHE	81.36	74	86.01	85.22	75.4	80.24	80.25	**87.39**	82.53
sonar	79.86	80.26	72.64	75.5	67.31	78.31	78.88	83.64	**90.38**
WarpAR10P	48.46	**74.62**	73.85	67.69	49.23	-	73.85	60.77	60
WarpPIE10P	93.33	95.71	95.71	95.71	95.24	-	94.76	95.24	**98.1**
Yale	57.46	62.9	57.46	61.1	55.07	-	40.07	62.32	**67.88**
Average	77.78	82.61	85.29	85.63	75.94	83.05	76.34	84.37	**88**

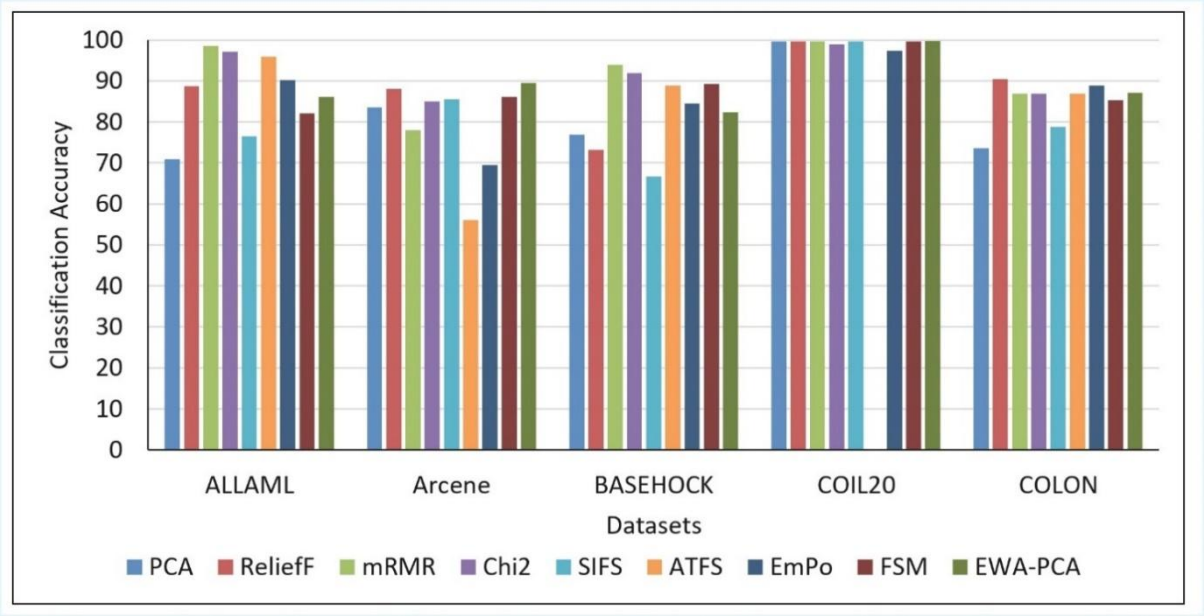

Fig 2: Comparison of CA of KNN Classifier Between Proposed EWA-PCA with Other Method on ALLAML, Arcene, BASEHOCK, COIL20, and COLON Datasets.

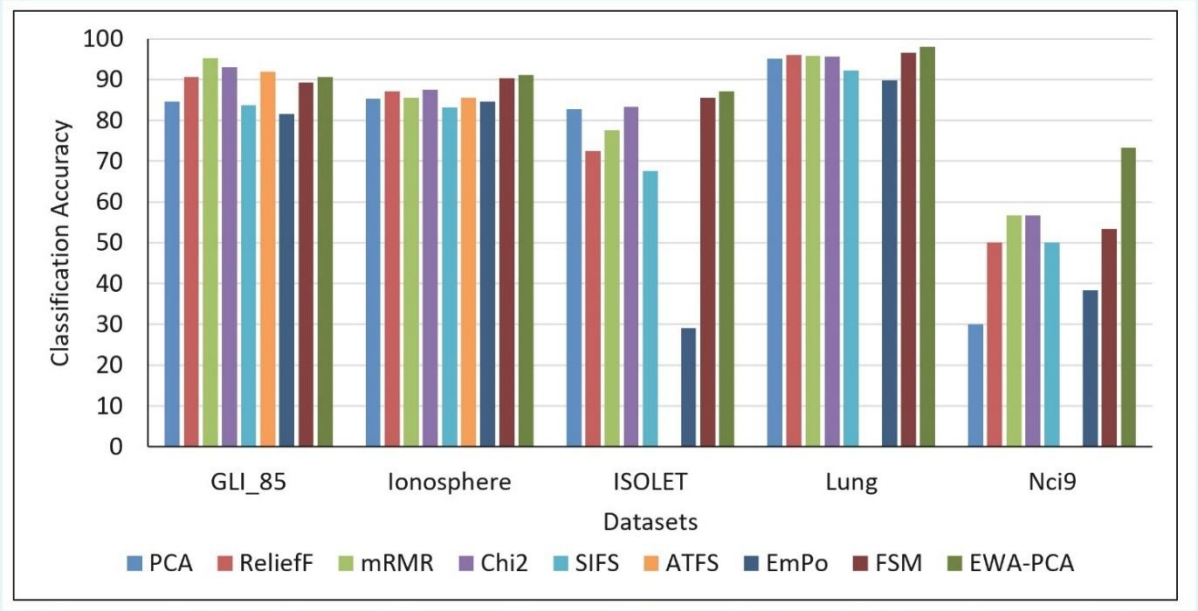

Fig 3: Comparison of CA of KNN Classifier Between Proposed EWA-PCA with Other Method on GLI_85, Ionosphere, ISOLET, Lung, and Nci9 Datasets.

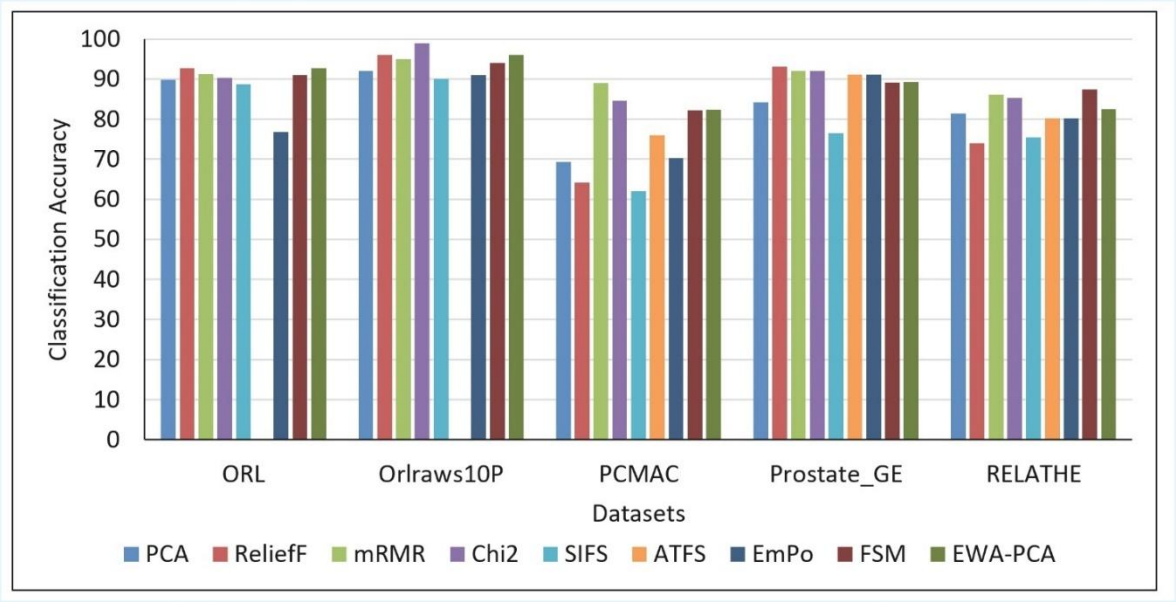

Fig 4: Comparison of CA of KNN Classifier Between Proposed EWA-PCA with Other Method on ORL, Orlraws10P, PCMAC, Prostate_GE, and RELATHE Datasets.

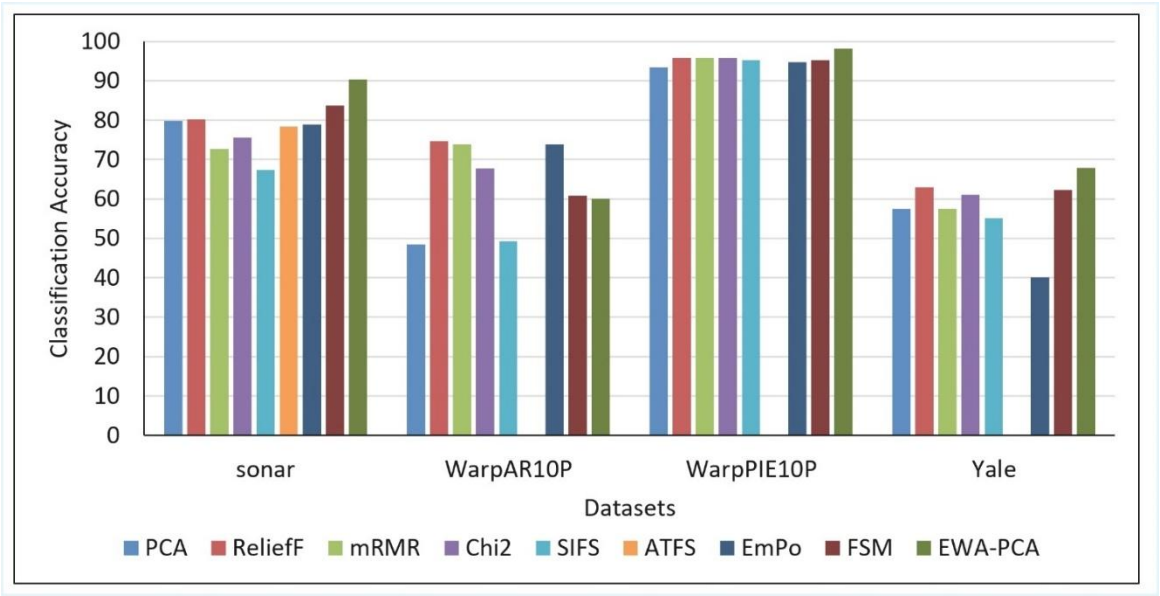

Fig 5: Comparison of CA of KNN Classifier Between Proposed EWA-PCA with Other Method on Sonar, WarpAR10P, WarpPIE10P, and Yale Datasets.

Table 7: Comparison of CA (In %) Between Proposed EWA-PCA with Eight Other FS Method using SVM classifier.

Datasets	PCA	ReliefF	mRMR	Chi2	SIFS	ATFS	EmPo	FSM	EWA-PCA
ALLAML	65.36	95.71	**98.57**	98.57	76.25	95.89	85.89	90.18	90.28
Arcene	82	**87.5**	83	84	86	56	68.5	87.5	83
BASEHOCK	95.38	83.4	93.98	92.68	66.03	86.15	95.99	97.54	**96.14**
COIL20	95.76	94.44	94.31	94.51	94.79	-	89.1	98.89	**99.93**
COLON	64.52	87.14	87.14	85.48	82.14	86.9	87.14	85.48	**90.32**
GLI_85	69.72	93.06	**94.16**	91.67	87.22	89.44	74.44	91.81	88.24
Ionosphere	91.73	93.17	89.45	90.89	85.18	92.31	85.75	93.73	**95.16**
ISOLET	95	94.74	94.55	95.38	95.06	-	40.32	94.81	**96.35**
Lung	93.62	68.5	68.5	76.86	68.5	-	69	88.67	**95.05**
Nci9	8.33	23.33	13.33	31.67	13.33	-	20	10	**25**

ORL	96	95	95.25	95.25	94.5	-	80.5	96.25	**96.75**
Orlraws10P	84	**99**	**99**	**99**	**99**	-	85	**99**	**99**
PCMAC	86.72	76.52	85.44	88.01	56.92	78.49	79.36	**92.69**	91.2
Prostate_GE	91.18	**94.09**	93.09	92.09	91.09	91.09	93.18	93.09	93.14
RELATHE	90.05	75.41	91.31	85.07	78.27	80.1	84.58	89.07	**91.17**
sonar	78.86	76.88	69.24	74.55	55.31	63.47	77.81	78.81	**87.02**
WarpAR10P	45.38	93.08	95.38	95.38	96.92	-	71.54	**100**	80.77
WarpPIE10P	99.52	99.05	98.57	98.1	98.57	-	92.86	**100**	**100**
Yale	72.61	76.25	78.71	73.27	70.18	-	47.79	**81.73**	76.97
Average	79.25	84.54	85.42	86.44	78.7	81.98	75.2	87.86	**88.16**

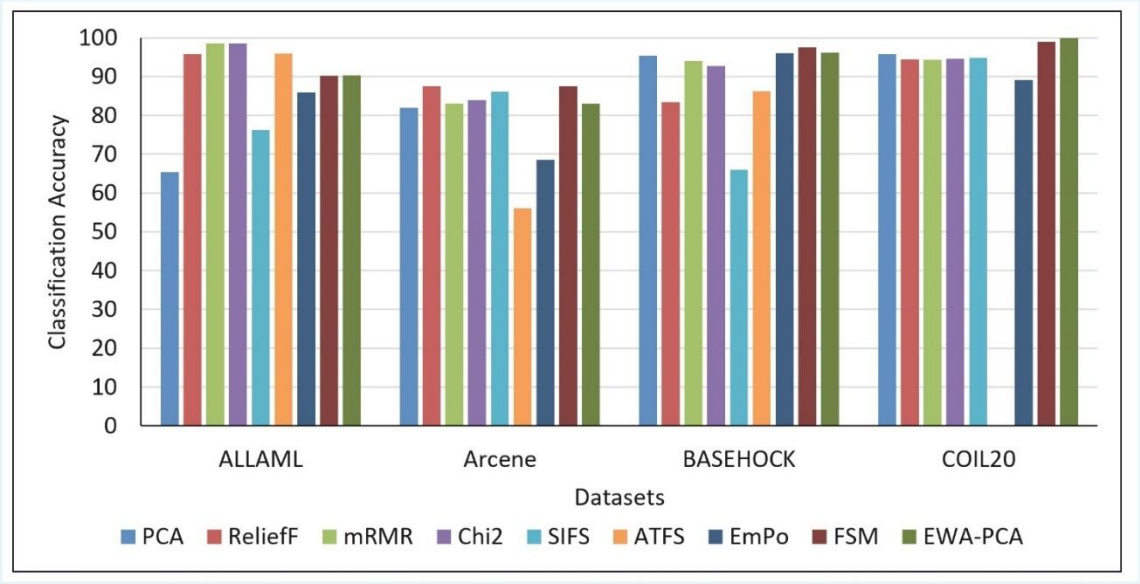

Fig 6: Comparison of CA of SVM Classifier Between Proposed EWA-PCA with Other Method on ALLAML, Arcene, BASEHOCK, COIL20, and COLON Datasets.

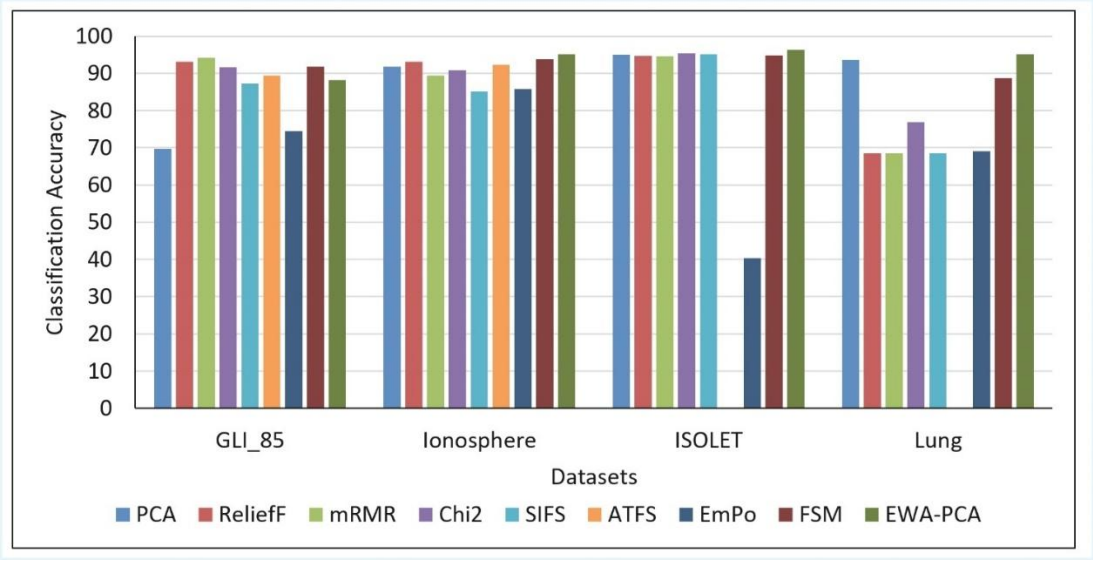

Fig 7: Comparison of CA of SVM classifier Between Proposed EWA-PCA with Other Method on GLI_85, Ionosphere, ISOLET, Lung, and Nci9 Datasets

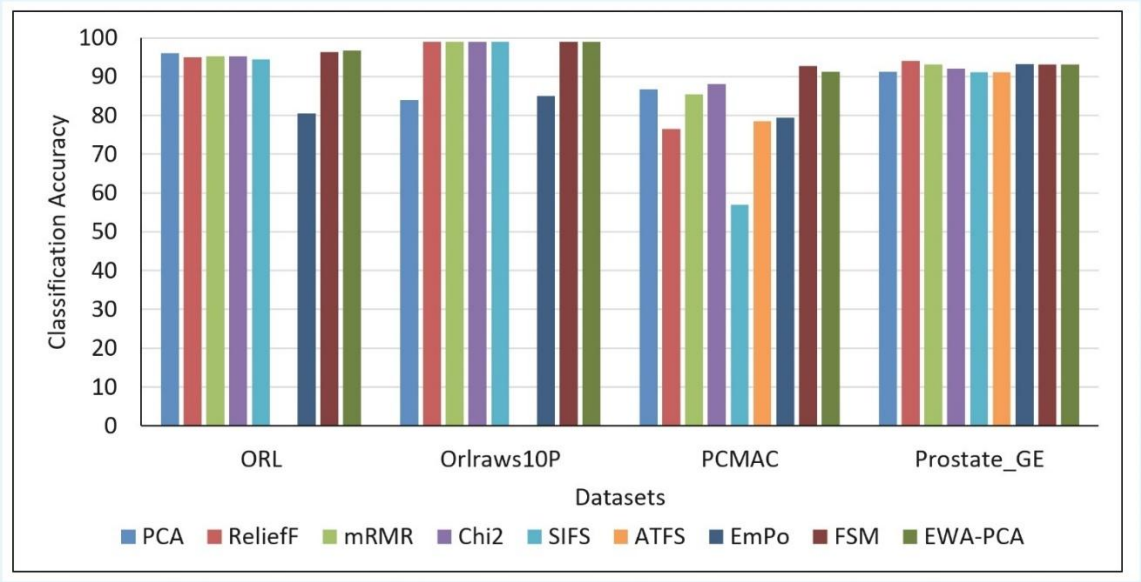

Fig 8: Comparison of CA of SVM Classifier Between Proposed EWA-PCA with Other Method on ORL, Orlraws10P, PCMAC, Prostate_GE, and RELATHE Datasets

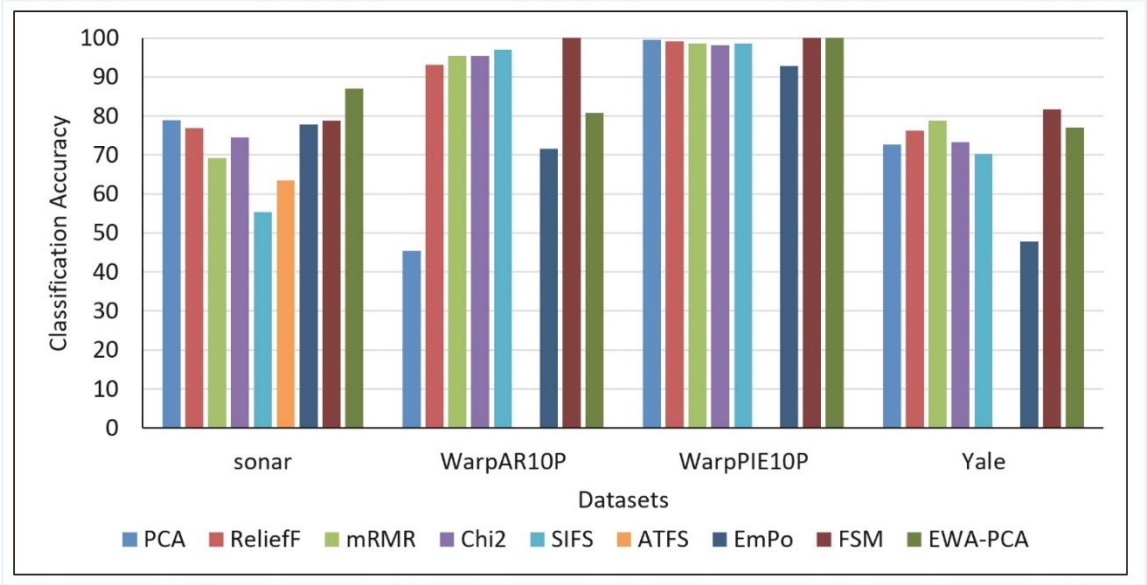

Fig 9: Comparison of CA of SVM Classifier Between Proposed EWA-PCA with Other Method on Sonar, WarpAR10P, WarpPIE10P, and Yale Datasets

Furthermore, compare the proposed EWA-PCA with eight other FS methods, namely: PCA, ReliefF, mRMR, Chi2, SIFS, ATFS, EmPo, and FSM. The ATFS method applied only binary dataset. Tables 6 and 7 provide a comparative analysis of the EWA-PCA against these algorithms using the KNN and SVM classifiers, respectively.

In Table 6, the findings clearly show that the proposed EWA-PCA provides higher CA on average (CA of 88%) across all datasets when compared to the other eight algorithms. This improvement in performance shows how well the features selected enhance KNN's ability for classification. Additionally, while the overall trend is positive, there are a few datasets where the KNN classifier experienced a slight drop in accuracy after FS. These slight decreases could be attributed to KNN's sensitivity to the local structure of the data, where neighborhood-based classification would have been impacted by the removal of certain features. However, despite these few cases, the proposed algorithm consistently outperforms or matches traditional methods in most scenarios, making it a reliable choice for enhancing KNN performance through effective feature reduction and selection. Similar to the KNN results, in Table 7, the selected feature sets generally yield a higher average CA (88.16%) when compared to the other eight FS algorithms using SVM, confirming the robustness of the proposed FS method. The average CA across all datasets shows noticeable improvement, indicating that SVM, with its margin-maximizing decision boundary, benefits significantly from the removal of irrelevant or redundant features. However, there are a few datasets where the SVM classifier exhibits a slight decrease in accuracy after FS using the proposed EWA-PCA. These cases may be due to the fact that some features, although redundant, contributed marginally to defining the optimal separating hyperplane. Nevertheless, the overall consistency in performance gains across the majority of datasets reaffirms the capability of the EWA-PCA to enhance SVM classification by focusing on the most informative features, leading to a more

generalizable and efficient learning model.

Figures 2 through 5 graphically compare the CA of the KNN classifier across 19 datasets, using the proposed EWA-PCA method versus eight alternative feature selection approaches. In Figure 2, which likely shows individual dataset performance (e.g., ALLAML, Arcene, BASEHOCK, COIL20, and COLON), EWA-PCA achieves the highest CA on the maximum datasets. Similarly, Figures 3, 4, and 5, which aggregate performance across the remaining datasets or show average CA, reveal that EWA-PCA maintains superior performance more consistently than any other method. Altogether, these visuals clearly demonstrate that the proposed EWA-PCA method delivers the highest classification accuracy with KNN on the majority of datasets, showcasing its robustness and effectiveness relative to existing feature-selection strategies.

Figures 6 to 9 present a graphical comparison of the CA of an SVM classifier across 19 datasets, contrasting the proposed EWA-PCA method with eight alternative feature-selection techniques. In Figure 6, which focuses on datasets such as ALLAML, Arcene, BASEHOCK, COIL20, and COLON, EWA-PCA achieves the highest CA in the majority of cases. The trend continues throughout Figures 7, 8, and 9, whether showing aggregated results across the remaining datasets or average CA EWA-PCA consistently outperforms all other methods. Collectively, these figures demonstrate that the proposed EWA-PCA method delivers the best SVM classification accuracy on most datasets, highlighting its robustness and superiority over existing feature-selection strategies.

Table 8: Comparison of the size of feature subsets

Datasets	Original	FSM Selected Feature for KNN	Proposed EWA-PCA Selected feature for KNN	FSM Selected Feature for SVM	Proposed EWA-PCA Selected feature for SVM
ALLAML	7129	375	59(3464)	387	59(3549)
Arcene	10000	477	135(5019)	551	135(5044)
BASEHOCK	4862	147	836(2407)	206	1032(4862)
COIL20	1024	120(512)	79(472)	170(922)	79(516)
COLON	2000	174	44(1003)	150	44(1015)
GLI_85	22283	1120	71(11202)	997	71(11082)
Ionosphere	34	14	10(12)	17	10(12)
ISOLET	617	132	115(308)	239(555)	157(617)
Lung	3312	662	134(1660)	632	140(1647)
Nci9	9712	50(4842)	50(4842)	50(4856)	50(4823)
ORL	1024	163(768)	100(520)	175(922)	98(496)
Orlraws10P	10304	3864	58(5168)	70(7213)	58(5795)
PCMAC	3289	171	162(1621)	174	481(3289)
Prostate_GE	5966	305	61(3005)	312	55(2934)
RELATHE	4322	231	632(2167)	289	736(3727)
sonar	60	9	18(29)	9	18(27)
WarpAR10P	2400	900	44(1148)	64(1800)	44(1194)
WarpPIE10P	2420	42(2178)	21(1200)	38(1815)	22(1167)
Yale	1024	84(512)	67(494)	78(512)	67(506)

NOTE: - It is important to note that the feature size following EWA-PCA dimensionality reduction is presented before the parentheses, whereas the feature size after EWA reduction to a single subset is specified within the parentheses.

Table 8 presents a comparison between the average size of feature subsets selected by EWA-PCA, FSM, and the original features. The findings indicate that EWA-PCA significantly decreases feature dimensionality by one to two orders of magnitude. In the arcene dataset, initially comprising 10000 features, EWA-PCA effectively minimized the subset size to 135 for KNN and 135 for SVM. Even when EWA simplifies selecting a single feature subset, it still achieves a reduction in the number of features by about 50%.

Based on the comparative analyses presented in Tables 4 and 5 using KNN and SVM classifiers, it is evident that the proposed EWA-PCA FS method consistently delivers superior or competitive performance across a wide range of datasets when compared to eight well-established FS techniques. While minor drops in accuracy were observed in a few individual datasets, the overall trend clearly favors the EWA-PCA, with notable improvements in average CA. The EWA method effectively balances the trade-off between removing redundant features and preserving the most informative ones, thereby enhancing the predictive power of both distance-based and margin-based classifiers.

6. CONCLUSION AND FUTURE WORK

In the proposed EWA-PCA approach, feature selection is performed in two stages to effectively reduce dimensionality while maintaining high classification performance. First, the Earthworm Optimization Algorithm (EWA) is applied to select

the most relevant features from the high-dimensional dataset. EWA identifies a subset of informative features that contribute significantly to classification accuracy while minimizing the overall feature count. In the second stage, Principal Component Analysis (PCA) is applied to the EWA-selected features to further reduce dimensionality by transforming the data into a lower-dimensional space. This step helps eliminate redundancy and enhance computational efficiency. The combination of EWA and PCA leads to a compact, high-quality feature representation that improves classification performance using models such as SVM and KNN.

For future research, the framework may be extended in several directions. One promising line of work is its integration with deep learning architectures to enhance feature representation in complex data domains. Another avenue is the application of the framework to real-time data streams and multi-class problems, where adaptability and speed are critical. Furthermore, the scalability, generalizability, and domain-specific performance of EWA-PCA can be further improved through the design of adaptive mechanisms and hybrid models.

7. REFERENCES

[1] Saxena, A. K., Dubey, V. K., and Wang, J. 2017. Hybrid feature selection methods for high-dimensional multi-class datasets. International Journal of Data Mining, Modelling and Management, 9(4), 315-339.

[2] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. 2016. Deep learning. 1(2).

[3] Theng, D., and Bhoyar, K. K. 2024. Feature selection techniques for machine learning: a survey of more than two decades of research. Knowledge and Information Systems, 66(3), 1575-1637.

[4] Lin, K. L., Lin, C. Y., Huang, C. D., Chang, H. M., Yang, C. Y., Lin, C. T., ... and Hsu, D. F. 2007. Feature selection and combination criteria for improving accuracy in protein structure prediction. IEEE Transactions on Nanobioscience, 6(2), 186-196.

[5] Saxena, A., Kothari, M., and Pandey, N. 2009. Evolutionary approach to dimensionality reduction. In Encyclopedia of Data Warehousing and Mining, Second Edition. 810-816.

[6] Ruano-Ordás, D. 2024. Machine learning-based feature extraction and selection. Applied Sciences, 14(15), 6567.

[7] Rahmat, F., Zulkafli, Z., Ishak, A. J., Abdul Rahman, R. Z., Stercke, S. D., Buytaert, W., ... and Ismail, M. 2024. Supervised feature selection using principal component analysis. Knowledge and Information Systems, 66(3), 1955-1995.

[8] Sánchez-Maroño, N., Alonso-Betanzos, A., and Tombilla-Sanromán, M. 2007. Filter methods for feature selection–a comparative study. In International conference on intelligent data engineering and automated learning. 178-187

[9] Patel, D., Saxena, A. K., Laha, S., and Ansari, G. M. 2022. A novel scheme for feature selection using filter approach. In 2022 7th International Conference on Computing, Communication and Security (ICCCS). 1-4.

[10] Dubey, V. K., and Saxena, A. K. 2016. Cosine similarity based filter technique for feature selection. In 2016 International Conference on Control, Computing, Communication and Materials (ICCCCM). 1-6.

[11] Batina, L., Gierlichs, B., Prouff, E., Rivain, M., Standaert, F. X., and Veyrat-Charvillon, N. 2011. Mutual information analysis: a comprehensive study. Journal of Cryptology, 24(2), 269-291.

[12] Patel, D., Saxena, A., and Wang, J. 2024. A machine learning-based wrapper method for feature selection. International Journal of Data Warehousing and Mining (IJDWM), 20(1), 1-33.

[13] Liu, H., Zhou, M., and Liu, Q. 2019. An embedded feature selection method for imbalanced data classification. IEEE/CAA Journal of Automatica Sinica, 6(3), 703-715.

[14] Saxena, A. K., and Dubey, V. K. 2015. A survey on feature selection algorithms. International Journal on Recent and Innovation Trends in Computing and Communication, 3(4), 1895-1899.

[15] Wang, G. G., Deb, S., and Coelho, L. D. S. 2018. Earthworm optimisation algorithm: a bio-inspired metaheuristic algorithm for global optimisation problems. International journal of bio-inspired computation, 12(1), 1-22.

[16] Robnik-Šikonja, M., and Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1), 23-69.

[17] Ding, C., and Peng, H. 2005. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[18] Eskandari, S., and Akbas, E. 2017. Supervised infinite feature selection. arXiv preprint arXiv:1704.02665.

[19] Abasabadi, S., Nematzadeh, H., Motameni, H., and Akbari, E. 2021. Automatic ensemble feature selection using fast non-dominated sorting. Information Systems, 100, 101760.

[20] Yin, Z., Yang, X., Wang, P., Yu, H., and Qian, Y. 2023. Ensemble selector mixed with pareto optimality to feature reduction. Applied Soft Computing, 148, 110877.

[21] McHugh, M. L. 2013. The chi-square test of independence. Biochemia medica, 23(2), 143-149.

[22] Liu, J., Li, D., Shan, W., and Liu, S. 2024. A feature selection method based on multiple feature subsets extraction and result fusion for improving classification performance. Applied Soft Computing, 150, 111018.

[23] Deng, Z., Zhu, X., Cheng, D., Zong, M., and Zhang, S. 2016. Efficient kNN classification algorithm for big data. Neurocomputing, 195, 143-148.

[24] Saxena, A., and Wang, J. 2012. Dimensionality reduction with unsupervised feature selection and applying non-Euclidean norms for classification accuracy. In Exploring Advances in Interdisciplinary Data Mining and Analytics: New Trends. 91-109.

[25] Wang, Q. 2022. Support vector machine algorithm in machine learning. In 2022 IEEE international conference on artificial intelligence and computer applications (ICAICA). 750-756.