The Scientific Temper (2024) Vol. 15 (spl): 86-96

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.11

https://scientifictemper.com/

RESEARCH ARTICLE

Cultural algorithm based principal component analysis (CA-PCA) approach for handling high dimensional data

G. Chitra, Hari Ganesh S.*

Abstract

The exponential growth of high-dimensional data in various domains, such as healthcare, finance, and image processing, presents significant challenges for efficient analysis and predictive modeling. Dimensionality reduction is a key technique to address these challenges, mitigating the curse of dimensionality while preserving the most relevant information. This paper proposes an optimization-based dimensionality reduction approach that integrates principal component analysis (PCA) with cultural algorithm (CA) optimization to enhance the handling of high-dimensional datasets. PCA is employed to transform the data by extracting principal components that capture the maximum variance. However, the selection of an optimal subset of components remains crucial for maintaining model accuracy and computational efficiency. To this end, the cultural algorithm is leveraged to optimize the selection of the most informative principal components by mimicking the evolutionary process of knowledge acquisition in a cultural framework. The proposed approach is validated through experiments on various high-dimensional datasets, demonstrating its superiority in reducing data dimensionality while maintaining high classification accuracy and reducing computational costs. The results highlight the effectiveness of combining PCA with cultural algorithm optimization for dimensionality reduction, paving the way for its application in large-scale real-world problems.

Keywords: Dimensionality reduction, Principal component analysis, Cultural algorithm, Healthcare domain.

Introduction

The healthcare domain encompasses a vast and complex ecosystem dedicated to maintaining and improving human health. It includes a wide range of stakeholders, such as healthcare providers (hospitals, clinics, and physicians), patients, pharmaceutical companies, government regulatory bodies, insurance companies, and technology vendors. This domain is driven by the mission to diagnose, treat, prevent, and manage diseases and medical conditions while improving the quality of life for individuals and communities. As healthcare continues to evolve, it is becoming more data-driven, personalized, and patient-centric, which requires the integration of advanced technologies, including data analytics, artificial intelligence (AI), machine learning (ML), and the Internet of Things (IoT), Ayesha, S., Hanif, M. K., & Talib, R. (2021), Hasan, B. M. S., & Abdulazeez, A. M. (2021).

In the modern healthcare landscape, vast amounts of data are generated daily, from electronic health records (EHRs) and wearable health devices to medical imaging and genomic sequencing. This data is essential for making informed decisions regarding patient care, diagnosis, and treatment, as well as improving operational efficiency in healthcare systems. Data-driven approaches have revolutionized healthcare by enabling personalized medicine, predictive analytics, and population health management. However, with the exponential increase in data volume and complexity, handling and analyzing high-dimensional healthcare data has become one of the greatest challenges, Ray, P., Reddy, S. S., & Banerjee, T. (2021), Patra, S. S., *et al.* (2021), Tripathy, B. K., Sundareswaran, A., & Ghela, S. (2021).

High-dimensional data in healthcare refers to datasets with numerous variables or features, often in the range of thousands to millions. For example, in genomics, each patient's DNA can be represented by millions of genetic markers, and in medical imaging, each scan can contain hundreds of thousands of pixels, each representing

PG & Research Department of Computer science, H.H. The Rajah's College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Pudukkottai, Tamilnadu, India.

***Corresponding Author:** Hari Ganesh S., PG & Research Department of Computer science, H.H. The Rajah's College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Pudukkottai, Tamilnadu, India., E-Mail: hariganesh17@gmail.com

How to cite this article: Chitra, G., Ganesh, H. S. (2024). Cultural algorithm based principal component analysis (CA-PCA) approach for handling high dimensional data. The Scientific Temper, **15**(spl):86-96.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.11

Source of support: Nil

Conflict of interest: None.

Received: 20/09/2024 **Accepted:** 30/09/2024 **Published**: 16/10/2024

important information. Such data, while rich in information, poses significant challenges for traditional data analysis and machine learning methods, which can suffer from the "curse of dimensionality." The curse of dimensionality refers to the phenomenon where the performance of algorithms deteriorates as the number of features increases, leading to increased computational complexity, noise, and overfitting in models, Alhassan, A. M., & Wan Zainon, W. M. N. (2021), Islam, M. T., & Xing, L. (2021), Nanga, S., *et al.* (2021), Poornappriya, T. S., & Durairaj, M. (2019), Durairaj, M., & Poornappriya, T. S. (2020).

Healthcare data is highly diverse and includes various types of structured and unstructured data:

Electronic Health Records (EHRs)

Digital versions of patient medical histories, including diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and lab test results. EHRs are a rich source of structured and semi-structured data that can be mined for insights into patient outcomes, disease progression, and treatment effectiveness.

Medical Imaging

This includes X-rays, MRIs, CT scans, and ultrasound images, which are used to diagnose and monitor diseases. The high-dimensional nature of medical imaging data makes it a critical area for dimensionality reduction and image processing techniques.

Genomic Data

Genomics focuses on the study of an individual's DNA and its impact on health. High-throughput technologies like next-generation sequencing (NGS) generate massive datasets that contain thousands to millions of genetic variants, making dimensionality reduction essential for identifying key markers associated with diseases.

Wearable Devices and IoT Data

The proliferation of wearable devices, such as fitness trackers and smartwatches, has led to continuous streams of real-time health data, including heart rate, blood pressure, and physical activity levels. This real-time data is critical for monitoring chronic conditions and enabling proactive healthcare interventions.

Clinical Notes

Unstructured text data from physician notes, discharge summaries, and patient interactions are often filled with valuable insights but require advanced natural language processing (NLP) techniques to extract meaningful information.

Principal Component Analysis Approach

Principal Component Analysis (PCA) is one of the most widely used techniques for dimensionality reduction in data science, machine learning, and statistics. It is a statistical method that transforms high-dimensional data into a lower-dimensional form while preserving as much variability (information) as possible. PCA achieves this by projecting the data onto a new set of orthogonal axes, known as principal components, which are ordered by the amount of variance they capture from the original dataset, Hasan, B. M. S., & Abdulazeez, A. M. (2021), Kostick, K. M., *et al.* (2021).

Dimensionality Reduction

PCA aims to reduce the number of variables (features) in a dataset while maintaining as much information as possible. This is especially useful in high-dimensional data, where many features may be correlated, redundant, or irrelevant.

Variance Preservation

PCA ensures that the principal components capture the maximum variance from the original data. The first principal component captures the most variance. The second captures the next most, and so on.

Data Visualization

PCA is often used for data visualization, especially in cases where the original dataset has more than three dimensions. By reducing the data to two or three principal components, one can create 2D or 3D plots that represent complex data structures.

PCA is based on linear algebra and involves finding the eigenvalues and eigenvectors of a data covariance matrix. The eigenvectors represent the principal components (new axes) onto which the original data will be projected, and the eigenvalues indicate the amount of variance captured by each principal component.

Step by Step Procedure for PCA

Step 1: Standardization

Given a dataset X of size n×d times dn×d, where n is the number of samples and d is the number of features, standardize the data if the features are measured on different scales. For each feature x_j, subtract the mean and divide by the standard deviation: $X' = \frac{X-\mu_j}{\sigma_j}$ where X' is the standardized data matrix, μ_j is the mean of feature j, σ_j is the standard deviation of feature j. This step ensures that all features contribute equally to the analysis.

Step 2: Compute the Covariance Matrix

Once the data is standardized, calculate the covariance matrix of the dataset X. The covariance matrix measures how different features in the data vary together. It is computed as: $\Sigma = \frac{1}{n}X^T X$ Where X^T is the transpose of the matrix X, Σ is the d X d covariance matrix. Each element Σ_{ij} of the covariance matrix represents the covariance between features i and j. If features are positively correlated Σ_{ij} will be positive, and if they are negatively correlated, Σ_{ij} will be negative.

Step 3: Eigenvalue and Eigenvector Decomposition

The next step is to perform eigenvalue decomposition on the covariance matrix Σ. This decomposition allows us to find the eigenvectors (principal components) and the corresponding eigenvalues (variance explained by each component). Solve the following equation for the eigenvalue λ and eigenvectors v: Σv=λv, where v represents the eigenvectors (principal components), λ represents the eigenvalues.

Eigenvalues λ_i represent the amount of variance explained by the corresponding eigenvector (principal component) $\mathbf{v}_i$. The larger the eigenvalue, the more variance is captured by that principal component.

Step 4: Sort Eigenvalues and Eigenvectors

Sort the eigenvalues in descending order and rearrange the eigenvectors accordingly. This ordering helps prioritize the principal components that capture the most variance. Let $\lambda_1 \geq \lambda_{i2} \geq \cdots \geq \lambda_d$, where λ_1 is the largest eigenvalue, corresponding to the first principal component, and so on.

Step 5: Select Principal Components

Decide how many principal components to retain based on the explained variance. The explained variance for each principal component i is given by: Explained Variance Ratio = $\frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}$. You may retain the top kkk principal components such that they capture a significant portion (e.g., 90% or 95%) of the total variance: $\sum_{i=1}^{k} \lambda_i \ \geq$ desired percentage of variance. The smaller the number of principal components, the more the data is compressed, but the less information it retains.

Step 6: Project the Data onto the Principal Components

Once you have selected the principal components (eigenvectors), project the original data X onto the new set of axes defined by the eigenvectors. Let $Z = XV_k$, where Z is the transformed dataset (with reduced dimensions), V_k is a d X k matrix, where each column is one of the top k eigenvectors, Z is of size n×k, representing the dataset in the new k-dimensional subspace.

Cultural Algorithm for Optimization

Cultural Algorithm (CA) is an evolutionary optimization technique inspired by the concept of cultural evolution in human societies. Unlike traditional genetic algorithms or evolutionary strategies, where the search for optimal solutions is driven purely by biological evolution, cultural algorithms utilize both individual experiences and shared cultural knowledge to guide the search process. This incorporation of a knowledge-sharing mechanism helps to improve the convergence speed and the quality of solutions in complex optimization problems, Coello, C. A. C., & Castillo Tapia, M. G. (2021), Meng, X. (2021).

Cultural Algorithms are modeled on two levels of evolution:

Population Space

This represents the traditional evolutionary algorithms (like Genetic Algorithms) where individual solutions evolve through variation and selection. The population space mimics biological evolution, where individuals are modified by genetic operators such as selection, crossover, and mutation.

Belief Space

This represents cultural evolution, where individuals share and accumulate knowledge over generations. The belief space serves as a repository of knowledge or rules that influence the evolutionary process. It is used to guide individuals in the population space toward more promising areas of the search space.

Cultural algorithms work by maintaining and updating both the population space (individual solutions) and the belief space (shared knowledge) throughout the optimization process. These two spaces interact to accelerate the search for optimal solutions.

Components of Cultural Algorithms

Population Space

- This is a collection of candidate solutions (individuals) to the optimization problem.
- Individuals in the population are evaluated based on a fitness function that reflects how well they solve the problem.
- Typical genetic operators, such as selection, crossover, and mutation, are applied to generate new individuals and evolve the population.

Belief Space

- The belief space represents accumulated knowledge that influences the evolutionary process.
- Knowledge in the belief space is structured in different categories such as norms, situational knowledge, historical knowledge, topographical knowledge, etc.
- The belief space is updated based on the best-performing individuals in the population space. It provides guidance to the population in future generations.

Communication Protocol

Acceptance Function

The acceptance function determines which individuals from the population space can influence the belief space. Typically, high-performing individuals contribute their knowledge to the belief space, improving the optimization process in subsequent generations.

Influence Function

The influence function determines how the belief space affects the population space. This function guides the search process by shaping or modifying the evolutionary process (e.g., adjusting mutation rates or constraining crossover).

Working of Cultural Algorithms

The basic steps of a Cultural Algorithm can be summarized as follows:

Initialization

- Generate an initial population of candidate solutions randomly.
- Initialize an empty belief space, or populate it with preliminary knowledge if available.

Evaluation

- Evaluate the fitness of each individual in the population space using the fitness function.

Update Belief Space

- Select individuals (based on their fitness) to contribute to the belief space.
- Update the belief space based on the knowledge extracted from the selected individuals (e.g., ranges of parameter values or successful solution strategies).

Influence Population

- Use the belief space to guide the creation of new individuals.
- Apply the influence function to modify the genetic operators or constrain the search to regions suggested by the belief space.

Evolution in Population Space

- Apply evolutionary operators (selection, crossover, mutation) to generate a new population of individuals.
- Repeat the evaluation, update, and influence steps for a set number of generations or until a stopping criterion is met.

Convergence

- The algorithm terminates when a satisfactory solution is found, or a maximum number of generations is reached.

Structure of the Belief Space

The belief space in Cultural Algorithms is divided into different components, each representing a different form of knowledge. These components guide the population evolution by influencing the genetic operations and the search space exploration. Some common categories of knowledge in the belief space are:

Normative Knowledge

This component represents the norms or acceptable ranges for solution parameters. For example, it may define upper and lower bounds for certain variables, restricting the search space based on past successful solutions.

Situational Knowledge

This contains specific experiences or situations that have led to successful outcomes in previous generations. It helps guide the search process toward promising regions of the search space.

Domain-Specific Knowledge

This includes any specialized knowledge relevant to the specific optimization problem being solved. For example, in healthcare optimization it might include knowledge about patient behavior or treatment success rates.

Topographical Knowledge

This refers to information about the fitness landscape of the problem, such as regions of high or low fitness. It can help the algorithm avoid local optima and search more effectively for the global optimum.

Historical Knowledge

This is the accumulated knowledge of the best solutions and strategies used in previous generations. It helps retain important information that can guide the algorithm in future generations.

Proposed Cultural Algorithm Based Principal Component Analysis (Ca-Pca) Approach For Handling High Dimensional Data

In this optimization-based dimensionality reduction approach, principal component analysis (PCA) is used to reduce the dimensionality of high-dimensional data, while the cultural algorithm (CA) is applied to optimize the selection of principal components, ensuring that the reduced feature set maintains maximum variance and performance.

Step 1: Data Preprocessing

- *Step 1.1: Data Collection*

Gather high-dimensional data relevant to the problem domain, ensuring the dataset contains all necessary features.

- *Step 1.2: Data Cleaning*

Handle missing values, remove noise, and resolve inconsistencies. Normalize or standardize the data (e.g., Z-score normalization) to ensure all features have comparable scales.

Step 2: Apply Principal Component Analysis (PCA)

- *Step 2.1: Standardize the Dataset*

Standardize the data matrix such that each feature has zero mean and unit variance. This ensures that PCA is not biased by features with different scales.

- *Step 2.2: Compute the Covariance Matrix*

Calculate the covariance matrix of the standardized data. This matrix captures the relationships between different features in the dataset.

- *Step 2.3: Eigenvalue Decomposition*

Perform eigenvalue decomposition on the covariance matrix.

- *Step 2.4: Rank Principal Components*

Rank the eigenvalues in descending order. The eigenvalue represents the amount of variance explained by the principal component (PC). The corresponding eigenvectors are the directions of the PCs.

- *Step 2.5: Retain Top Principal Components*

Select the top k principal components that account for the majority of the variance. The number of components ks is determined by a cumulative variance threshold, (e.g., 90%)

Step 3: Initialize the Cultural Algorithm (CA)

- *Step 3.1: Population Initialization*
- Initialize a population of candidate solutions. Each candidate represents a potential subset of principal components selected from the PCA output.
- Each individual in the population is a binary vector, where a value of 1 indicates that the corresponding principal component is selected, and 0 indicates it is not selected.

- *Step 3.2: Belief Space Initialization*

Initialize the belief space, which stores shared knowledge and guides the optimization process. The belief space contains:

- Normative knowledge: Ranges for the number of principal components to select.
- Situational knowledge: Knowledge about successful combinations of principal components.

Step 4: Fitness Evaluation

- *Step 4.1: Define Fitness Function*

Design a fitness function to evaluate the performance of each candidate solution (subset of principal components). The fitness function should measure:

- Classification/Prediction Performance: Use a supervised learning algorithm (e.g., Support Vector Machine, Logistic Regression) to classify or predict based on the selected components.
- Dimensionality Reduction: Reward solutions with fewer components, ensuring effective dimensionality reduction without compromising on accuracy. Where α and β are weighting factors to balance accuracy and reduction.

$$Fitness = \alpha . Accuracy - \beta . Number\ of\ Components$$

- *Step 4.2: Evaluate Population*

For each individual in the population, select the subset of principal components and evaluate its fitness using the fitness function.

Step 5: Update the Belief Space

- *Step 5.1: Acceptance Function*

Use the acceptance function to select high-performing individuals from the population to update the belief space. Typically, the top-performing individuals contribute their knowledge to the belief space.

- *Step 5.2: Update Normative Knowledge*

Update the normative knowledge (e.g., upper and lower bounds for the number of components) based on successful candidates: B(t+1)normative = [min (components), max (components)]

- *Step 5.3: Update Situational Knowledge*

Update situational knowledge by identifying successful combinations of principal components that frequently appear in high-performing solutions. This knowledge can guide future generations.

Step 6: Generate New Population

- *Step 6.1: Apply Influence Function*

Use the belief space to influence the generation of new solutions. The influence function modifies the genetic operators or constrains the search space. For instance:

- Use normative knowledge to limit the range of principal components selected.
- Use situational knowledge to encourage or discourage certain combinations of components.

- *Step 6.2: Apply Genetic Operators*

Perform evolutionary operations such as

- Selection: Select individuals based on their fitness scores to create a mating pool.
- Crossover: Combine individuals to generate new solutions by exchanging components.
- Mutation: Randomly modify individual components to explore new regions of the search space.

Step 7: Convergence and Termination

- *Step 7.1: Convergence Check*

Monitor the optimization process for convergence. The algorithm terminates when:

- A satisfactory solution is found (i.e., a subset of principal components that achieves high performance and substantial dimensionality reduction).
- The fitness improvement between generations falls below a threshold.
- A maximum number of generations is reached.

- *Step 7.2: Final Solution*

The best-performing individual at the end of the optimization process is selected as the final solution. This individual represents the optimal subset of principal components for dimensionality reduction.

Result And Discussion

Performance Metrics

The performance of the proposed CA-PCA approach is evaluated with other optimization techniques like genetic algorithm, particle swarm optimization (PSO), artificial bee colony (ABC) and ant colony optimization (ACO) with PCA.

The following table metrics are considered to evaluate the performance. Table 1 depicts the performance metrics used in this paper.

To evaluate the performance of the Proposed PCA-CA approach for dimensionality reduction, the healthcare datasets that are publicly available are considered. The cervical cancer, lung cancer and dermatology datasets are considered to evaluate the proposed dimensionality reduction approach. The classification techniques like random forest (RF), K-nearest neighbor (KNN), gradient boosting tree (GBT), neural network (NN), and Naïve Bayes (NB) are considered UCI machine learning repositories (n.d.), cervical cancer risk factors, lung cancer, dermatology.

Classification Accuracy

Table 2 depicts the classification accuracy (in %) obtained by the Proposed and existing Optimization techniques like a genetic algorithm (GA), PSO, ABC, and ACO for cervical cancer dataset, lung cancer dataset and dermatology dataset using RF, KNN, GBT, ANN, and NB classifiers.

The classification accuracy for various datasets (Cervical Cancer, Lung Cancer, and Dermatology) using different optimization techniques (GA, PSO, ABC, ACO) and classifiers (random forest (RF), K-nearest neighbors (KNN), gradient boosting trees (GBT), artificial neural networks (ANN), and Naive Bayes (NB)) is presented in Table 2.

Cervical Cancer Dataset

- The original dataset yielded the lowest accuracies across all classifiers, ranging from 41.65 (NB) to 48.32% (GBT).
- Optimization techniques significantly improved the classification accuracy, with the ACO technique achieving the highest accuracy (72.87% for GBT) among the existing methods.
- The proposed CA-PCA method outperformed all existing techniques, achieving impressive accuracies of 93.55 (RF) and 95.06% (GBT).

Lung Cancer Dataset

- Similar to the cervical cancer dataset, the original dataset had low accuracy rates, with NB achieving the lowest at 41.75%.
- ACO again yielded the highest accuracy for existing methods (72.59% for GBT).
- The proposed CA-PCA method showed substantial improvements, with accuracies of 93.46 (RF) and 94.91% (GBT), indicating its effectiveness in enhancing classification performance.

Table 1: Performance metrics

Metrics	*Equation*
Accuracy	TP+TN/TP+FN+TN+FP
True positive rate (TPR) (Sensitivity or Recall)	TP/TP+FN
False positive rate (FPR)	FP/FP+TN
Precision	TP/TP+FP
True negative rate (Specificity)	1- False Positive Rate (FPR)
Miss rate	1-True Positive Rate (TPR)

Dermatology Dataset

- The original dataset's accuracies were again the lowest, with NB showing 42.72% as the minimum.
- ACO produced the highest accuracy among existing methods, achieving 74.04% for both GBT and KNN.
- The proposed CA-PCA method achieved the highest accuracies across all classifiers, with 95.85% (KNN) and 95.15% (RF and GBT), demonstrating its superiority in classification tasks.

True Positive Rate

Table 3 depicts the true positive rate (in %) obtained by the proposed and existing optimization techniques like GA, PSO, ABC, and ACO for cervical cancer dataset, lung cancer

Table 2: Classification accuracy (in %) obtained by the proposed and existing optimization techniques like GA, PSO, ABC, and ACO for the cervical cancer dataset, lung cancer dataset and dermatology dataset using RF, KNN, GBT, ANN, and NB classifiers

Feature selection methods	*Classification accuracy (in %) – Cervical cancer dataset*				
	RF	*KNN*	*GBT*	*NN*	*NB*
Original dataset	43.099	46.44	48.32	42.98	41.65
GA	69.63	69.97	70.84	68.45	67.91
PSO	66.54	66.86	68.75	63.34	62.82
ABC	65.46	65.77	67.64	62.23	61.73
ACO	71.76	72.30	72.87	69.81	68.27
Proposed CA-PCA	93.55	94.86	95.06	78.92	79.45
	Classification accuracy (in %) – Lung cancer dataset				
Original dataset	43.97	44.98	48.32	42.86	41.75
GA	69.34	70.94	70.84	67.43	66.83
PSO	58.43	59.85	59.73	56.32	55.72
ABC	57.34	58.74	58.64	55.43	54.81
ACO	71.67	71.47	72.59	69.78	68.92
Proposed CA-PCA	93.46	94.09	94.91	80.58	78.21
	Classification accuracy (in %) – dermatology dataset				
Original dataset	44.93	45.81	50.16	43.82	42.72
GA	68.81	67.11	66.19	64.28	63.22
PSO	57.92	58.22	55.28	53.37	52.34
ABC	56.81	57.32	54.19	52.46	51.45
ACO	74.04	72.20	74.04	69.15	68.42
Proposed CA-PCA	95.15	95.85	95.15	82.57	81.98

dataset and dermatology dataset using RF, KNN, GBT, ANN, and NB classifiers.

Table 3 presents the true positive rate (TPR) achieved by various optimization techniques (GA, PSO, ABC, ACO) and classifiers (Random Forest (RF), K-nearest neighbors (KNN), gradient boosting trees (GBT), artificial neural networks (ANN), and Naive Bayes (NB)) across three datasets: Cervical cancer, lung cancer, and dermatology.

Cervical Cancer Dataset

- The original dataset recorded relatively low TPR values, with NB showing the lowest at 50.26% and KNN the highest at 52.94%.
- All optimization techniques improved TPR significantly, with the GA method achieving the highest TPR of 76.07% for RF.
- ACO also yielded notable results, reaching a maximum TPR of 76.37% for KNN.
- The proposed CA-PCA method demonstrated remarkable performance, attaining TPRs of 93.35% (RF) and 94.99% (KNN), highlighting its effectiveness in enhancing classification sensitivity.

Lung Cancer Dataset

- The original dataset showed low TPR values, with the highest at 52.76% for GBT and the lowest at 45.87% for NB.
- GA provided significant improvements, especially with KNN (75.50%) and GBT (74.45%).
- ACO outperformed other existing techniques, achieving a TPR of 82.3% for RF.
- The proposed CA-PCA method excelled with TPRs of 92.42% (RF) and 95.51% (GBT), indicating substantial enhancements in sensitivity across classifiers.

Dermatology Dataset

- The original dataset produced the lowest TPR values again, with NB at 46.22% and RF at 55.11%.
- GA yielded improved results, with TPRs around 67.35% for RF and 70.57% for GBT.
- ACO showed the highest TPR among existing methods, reaching 84.55% for RF.
- The proposed CA-PCA method achieved the highest TPR values, with 96.53% (RF) and 96.90% (KNN), demonstrating its superior ability to identify positive cases in the dataset.

Table 3: True positive rate (in %) obtained by the proposed and existing optimization techniques like GA, PSO, ABC, and ACO for cervical cancer dataset, lung cancer dataset and dermatology dataset using RF, KNN, GBT, ANN, and NB classifiers

Feature selection methods	*True positive rate (in %) – Cervical cancer dataset*				
	RF	*KNN*	*GBT*	*NN*	*NB*
Original dataset	52.61	52.94	52.80	51.78	50.26
GA	76.07	74.59	71.35	73.63	72.21
PSO	69.18	67.68	65.45	65.72	64.32
ABC	64.34	66.57	68.29	64.61	63.21
ACO	75.37	76.37	70.54	69.32	68.54
Proposed CA-PCA	93.35	94.99	94.97	78.15	77.73
	True positive rate (in %) – Lung cancer dataset				
Original dataset	51.26	47.68	52.76	46.35	45.87
GA	73.05	75.50	74.45	71.16	70.61
PSO	62.16	64.41	63.34	60.24	59.72
ABC	61.27	63.32	62.25	59.13	58.61
ACO	82.3	74.90	71.19	69.24	67.72
Proposed CA-PCA	92.42	94.51	95.51	79.96	77.45
	True positive rate (in %) – Dermatology dataset				
Original dataset	55.11	49.44	54.26	47.53	46.22
GA	67.35	69.42	70.57	65.46	64.53
PSO	56.43	58.31	60.46	54.55	53.64
ABC	55.65	57.53	59.68	53.73	52.86
ACO	84.55	80.57	81.74	78.66	77.65
Proposed CA-PCA	96.53	96.90	95.02	81.69	80.85

Table 4: False positive rate (in %) obtained by the proposed and existing optimization techniques like GA, PSO, ABC, and ACO for cervical cancer dataset, lung cancer dataset and dermatology dataset using RF, KNN, GBT, ANN, and NB classifiers

Feature Selection Methods	*False Positive Rate (in %) – Cervical Cancer Dataset*				
	RF	*KNN*	*GBT*	*NN*	*NB*
Original dataset	67.17	61.08	56.83	68.89	69.04
GA	35.62	34.77	29.73	36.47	37.15
PSO	46.53	45.66	40.82	47.82	48.26
ABC	47.42	46.75	41.71	48.73	49.37
ACO	32.18	32.8	24.22	35.47	36.02
Proposed CA-PCA	6.21	5.26	4.84	27.36	25.14
	False Positive Rate (in %) – Lung Cancer Dataset				
Original dataset	63.8	57.67	56.58	64.32	65.52
GA	35.31	33.75	32.87	36.22	37.64
PSO	44.42	42.84	43.78	47.35	46.53
ABC	45.53	43.75	44.69	48.43	47.44
ACO	31.91	32.31	25.60	33.42	34.54
Proposed CA-PCA	5.36	6.36	5.72	20.36	20.54
	False Positive Rate (in %) – Dermatology Dataset				
Original dataset	64.74	59.04	54.40	65.85	66.15
GA	28.79	35.32	38.08	39.19	40.43
PSO	37.88	36.43	39.19	40.28	41.34
ABC	38.06	37.65	40.32	41.40	42.56
ACO	35.76	36.21	34.56	37.32	38.65
Proposed CA-PCA	6.32	5.305	4.704	20.73	21.36

False Positive Rate

Table 4 depicts the false positive rate (in %) obtained by the proposed and existing optimization techniques like GA, PSO, ABC, and ACO for cervical cancer dataset, lung cancer dataset and dermatology dataset using RF, KNN, GBT, ANN, and NB classifiers.

Table 4 presents the false positive rate (FPR) obtained through various optimization techniques (GA, PSO, ABC, ACO) and classifiers (Random Forest (RF), K-Nearest Neighbors (KNN), gradient boosting trees (GBT), artificial neural networks (ANN), and Naive Bayes (NB)) across three datasets: Cervical cancer, lung cancer, and dermatology.

Cervical Cancer Dataset

- The original dataset recorded high FPR values, with RF at 67.17% and KNN at 61.08%.
- All optimization techniques successfully reduced FPR, with ACO achieving the lowest rate of 24.22% for GBT.
- The GA method also provided significant improvements, lowering FPR to 29.73% for GBT.
- The proposed CA-PCA method drastically reduced FPR, reaching only 4.84% for GBT and 5.26% for KNN, indicating a substantial enhancement in classification reliability.

Lung Cancer Dataset

- Similar to the cervical cancer dataset, the original dataset exhibited high FPR values, peaking at 65.52% for NB.
- GA effectively decreased FPR, particularly for KNN (33.75%) and GBT (32.87%).
- ACO also demonstrated notable results with an FPR of 25.60% for GBT.
- The proposed CA-PCA method yielded remarkable results, achieving FPRs as low as 5.36% for RF and 6.36% for KNN, showcasing its effectiveness in reducing false positives.

Dermatology Dataset

- The original dataset's FPR was also high, with NB at 66.15% and RF at 64.74%.
- GA provided significant improvements, lowering FPR to 28.79% for RF.
- ACO resulted in an FPR of 34.56% for GBT, while ABC and PSO recorded similar reductions.
- The proposed CA-PCA method achieved exceptionally low FPR values, reaching 4.704% for GBT and 5.305% for KNN, indicating a strong capability to minimize false positive classifications.

Precision

Table 5 depicts the Precision (in %) obtained by the Proposed and existing Optimization techniques like GA, PSO, ABC, and ACO for Cervical Cancer Dataset, Lung Cancer Dataset and Dermatology Dataset using RF, KNN, GBT, ANN, and NB classifiers.

Table 5 presents the Precision values achieved by various optimization techniques (GA, PSO, ABC, ACO) and classifiers (Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting Trees (GBT), Artificial Neural Networks (ANN), and Naive Bayes (NB)) across three datasets: Cervical Cancer, Lung Cancer, and Dermatology.

Cervical Cancer Dataset

- The original dataset showed relatively low precision values, with RF at 45.81% and NB at 43.66%.
- All optimization techniques improved precision significantly, with ACO achieving the highest precision of 78.97% for GBT.
- The GA method also yielded high precision rates, reaching 73.60% for GBT.
- The proposed CA-PCA method dramatically outperformed existing techniques, achieving precision rates of 94.30% (RF) and 95.50% (GBT), indicating its effectiveness in enhancing the reliability of positive classifications.

Lung Cancer Dataset

- The original dataset's precision values were low, with NB at 44.83% and RF at 46.11%.

Table 5: Precision (in %) obtained by the Proposed and existing Optimization techniques like GA, PSO, ABC, and ACO for Cervical Cancer Dataset, Lung Cancer Dataset and Dermatology Dataset using RF, KNN, GBT, ANN, and NB classifiers

Feature Selection Methods	*Precision (in %) – Cervical Cancer Dataset*				
	RF	*KNN*	*GBT*	*NN*	*NB*
Original dataset	45.81	49.01	51.72	44.54	43.66
GA	68.79	68.81	73.60	67.89	66.34
PSO	59.68	59.72	62.51	56.78	55.43
ABC	58.57	58.61	61.43	55.65	54.32
ACO	71.97	71.45	78.97	70.54	69.80
Proposed CA-PCA	94.30	95.16	95.50	82.46	78.91
	Precision (in %) – Lung Cancer Dataset				
Original dataset	46.11	52.34	50.84	45.96	44.83
GA	69.21	69.77	70.04	67.32	65.98
PSO	58.32	58.68	61.13	56.31	54.87
ABC	57.43	57.79	60.24	55.42	53.78
ACO	72.76	71.92	78.18	69.53	68.25
Proposed CA-PCA	95.11	94.16	94.17	81.74	80.63
	Precision (in %) – Dermatology Dataset				
Original dataset	44.75	52.80	52.67	43.86	42.57
GA	74.80	67.58	64.97	63.86	62.67
PSO	63.91	57.47	53.86	52.77	51.56
ABC	61.13	56.69	52.08	51.95	50.78
ACO	68.81	69.09	72.55	67.62	66.18
Proposed CA-PCA	94.24	95.11	95.56	82.21	81.59

- GA provided substantial improvements, achieving precision of 70.04% for GBT and 69.77% for KNN.
- ACO also demonstrated strong performance, reaching 78.18% for GBT.
- The proposed CA-PCA method achieved precision values of 95.11% (RF) and 94.17% (GBT), highlighting its effectiveness in improving classification accuracy.

Dermatology Dataset

- The original dataset exhibited low precision, with NB at 42.57% and RF at 44.75%.
- GA showed significant improvements, with precision rates of 74.80% for RF and 67.58% for KNN.
- ACO also provided substantial results, achieving 72.55% for GBT.
- The proposed CA-PCA method excelled with precision values of 94.24% (RF) and 95.56% (GBT), further demonstrating its superiority in classification tasks.

Specificity

Table 6 depicts the Specificity (in %) obtained by the Proposed and existing Optimization techniques like GA, PSO, ABC, and ACO for Cervical Cancer Dataset, Lung Cancer Dataset and Dermatology Dataset using RF, KNN, GBT, ANN, and NB classifiers.

Table 6: Specificity (in %) obtained by the Proposed and existing Optimization techniques like GA, PSO, ABC, and ACO for Cervical Cancer Dataset, Lung Cancer Dataset and Dermatology Dataset using RF, KNN, GBT, ANN, and NB classifiers

Feature Selection Methods	*Specificity (in %) – Cervical Cancer Dataset*				
	RF	*KNN*	*GBT*	*NN*	*NB*
Original dataset	32.83	38.92	43.17	31.11	30.96
GA	64.38	65.23	70.27	63.53	62.85
PSO	53.47	54.34	59.18	52.18	51.74
ABC	52.58	53.25	58.29	51.27	50.63
ACO	67.82	67.2	75.78	64.53	63.98
Proposed CA-PCA	93.79	94.74	95.16	72.64	74.86
	Specificity (in %) – Lung Cancer Dataset				
Original dataset	36.2	42.33	43.42	35.68	34.48
GA	64.69	66.25	67.13	63.78	62.36
PSO	55.58	57.16	56.22	52.65	53.47
ABC	54.47	56.25	55.31	51.57	52.56
ACO	68.09	67.69	74.4	66.58	65.46
Proposed CA-PCA	94.64	93.64	94.28	79.64	79.46
	Specificity (in %) – Dermatology Dataset				
Original dataset	35.26	40.96	45.6	34.15	33.85
GA	71.21	64.68	61.92	60.81	59.57
PSO	62.12	63.57	60.81	59.72	58.66
ABC	61.94	62.35	59.68	58.6	57.44
ACO	64.24	63.79	65.44	62.68	61.35
Proposed CA-PCA	93.68	94.695	95.296	79.27	78.64

Table 6 presents the Specificity values achieved by various optimization techniques (GA, PSO, ABC, ACO) and classifiers (Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting Trees (GBT), Artificial Neural Networks (ANN), and Naive Bayes (NB)) across three datasets: Cervical Cancer, Lung Cancer, and Dermatology.

Cervical Cancer Dataset

- The original dataset showed low specificity values, with RF at 32.83% and NB at 30.96%.
- All optimization techniques improved specificity significantly, with ACO achieving the highest specificity of 75.78% for GBT.
- The GA method also yielded strong results, with a specificity of 70.27% for GBT.
- The proposed CA-PCA method significantly outperformed existing techniques, achieving specificity values of 93.79% (RF) and 94.74% (KNN), indicating a robust ability to correctly identify true negatives.

Lung Cancer Dataset

- The original dataset recorded low specificity values, with RF at 36.2% and NB at 34.48%.
- GA demonstrated notable improvements, reaching a maximum specificity of 67.13% for GBT.
- ACO also provided strong performance, achieving a specificity of 74.4% for GBT.
- The proposed CA-PCA method excelled with specificity values of 94.64% (RF) and 93.64% (KNN), demonstrating its effectiveness in accurately identifying non-cancerous cases.
- Dermatology Dataset
- The original dataset exhibited low specificity, with RF at 35.26% and NB at 33.85%.
- GA significantly improved specificity, achieving 71.21% for RF.
- ACO recorded a maximum specificity of 65.44% for GBT.
- The proposed CA-PCA method achieved high specificity rates of 93.68% (RF) and 94.695% (KNN), underscoring its capacity to correctly identify true negatives in dermatological assessments.

Miss Rate

Table 7 depicts the Miss Rate (in %) obtained by the Proposed and existing Optimization techniques like GA, PSO, ABC, and ACO for Cervical Cancer Dataset, Lung Cancer Dataset and Dermatology Dataset using RF, KNN, GBT, ANN, and NB classifiers.

Table 7 presents the Miss Rate values obtained using various optimization techniques (GA, PSO, ABC, ACO) and classifiers (Random Forest (RF), K-Nearest Neighbors (KNN),

Table 7: Miss Rate (in %) obtained by the Proposed and existing Optimization techniques like GA, PSO, ABC, and ACO for Cervical Cancer Dataset, Lung Cancer Dataset and Dermatology Dataset using RF, KNN, GBT, ANN, and NB classifiers

Feature Selection Methods	*Miss Rate (in %) – Cervical Cancer Dataset*				
	RF	*KNN*	*GBT*	*NN*	*NB*
Original dataset	47.39	47.06	47.2	48.22	49.74
GA	23.93	25.41	28.65	26.37	27.79
PSO	30.82	32.32	34.55	34.28	35.68
ABC	35.66	33.43	31.71	35.39	36.79
ACO	24.63	23.63	29.46	30.68	31.46
Proposed CA-PCA	6.65	5.01	5.03	21.85	22.27
	Miss Rate (in %) – Lung Cancer Dataset				
Original dataset	48.74	52.32	47.24	53.65	54.13
GA	26.95	24.5	25.55	28.84	29.39
PSO	37.84	35.59	36.66	39.76	40.28
ABC	38.73	36.68	37.75	40.87	41.39
ACO	17.7	25.1	28.81	30.76	32.28
Proposed CA-PCA	7.58	5.49	4.49	20.04	22.55
	Miss Rate (in %) – Dermatology Dataset				
Original dataset	44.89	50.56	45.74	52.47	53.78
GA	32.65	30.58	29.43	34.54	35.47
PSO	43.57	41.69	39.54	45.45	46.36
ABC	44.35	42.47	40.32	46.27	47.14
ACO	15.45	19.43	18.26	21.34	22.35
Proposed CA-PCA	3.47	3.1	4.98	18.31	19.15

Gradient Boosting Trees (GBT), Artificial Neural Networks (ANN), and Naive Bayes (NB)) across three datasets: Cervical Cancer, Lung Cancer, and Dermatology.

Cervical Cancer Dataset

- The original dataset exhibited high miss rates, with RF at 47.39% and NB at 49.74%.
- All optimization techniques significantly reduced the miss rate, with ACO achieving a low miss rate of 24.63% for RF.
- GA also showed substantial improvement, reaching a miss rate of 23.93% for RF.
- The proposed CA-PCA method demonstrated exceptional performance, achieving a miss rate as low as 5.01% for KNN and 5.03% for GBT, indicating its effectiveness in minimizing misclassifications.

Lung Cancer Dataset

- The original dataset had a high miss rate, particularly with KNN at 52.32% and NB at 54.13%.
- GA effectively reduced the miss rate to 24.50% for KNN.
- ACO also demonstrated a notable decrease, achieving a miss rate of 17.70% for RF.
- The proposed CA-PCA method excelled with a miss rate of only 4.49% for GBT and 5.49% for KNN, highlighting its capability in enhancing classification accuracy.

Dermatology Dataset

- The original dataset showed high miss rates, with RF at 44.89% and NB at 53.78%.
- GA significantly lowered the miss rates, reaching 29.43% for GBT.
- ACO achieved a low miss rate of 15.45% for RF.
- The proposed CA-PCA method achieved the best results, with miss rates of 3.10% for KNN and 3.47% for RF, demonstrating its superior ability to classify correctly and minimize misclassifications.

Conclusion

The proposed Optimization-Based Dimensionality Reduction Approach utilizing Principal Component Analysis (PCA) combined with Cultural Algorithm (CA) Optimization effectively addresses the challenges associated with high-dimensional data. In today's data-driven landscape, where vast amounts of information can overwhelm traditional analysis methods, the need for efficient dimensionality reduction techniques is paramount.

This approach leverages PCA's ability to transform high-dimensional datasets into a lower-dimensional space while preserving essential variance, thereby facilitating easier data interpretation and analysis. By incorporating CA, the method not only selects the most relevant principal components but also optimizes their selection based on a fitness function that balances accuracy and dimensionality reduction. This dual mechanism enhances the robustness and adaptability of the model, ensuring that it can perform well across different datasets and applications.

The iterative process of evaluating and updating both the population and belief space in CA promotes the exploration of diverse solutions, ultimately leading to improved convergence and performance. The belief space, enriched with normative and situational knowledge, guides the optimization process effectively, fostering the selection of components that contribute to superior predictive accuracy.

The results from the evaluations across the Cervical Cancer, Lung Cancer, and Dermatology datasets highlight the effectiveness of the proposed CA-PCA method in enhancing classification performance when compared to existing optimization techniques such as GA, PSO, ABC, and ACO.

The proposed method consistently achieved superior results in key performance metrics, including classification accuracy, true positive rate, specificity, precision, and notably, the miss rate. For instance, in all three datasets, the CA-PCA method demonstrated remarkable capabilities in correctly identifying cases while minimizing misclassifications, evidenced by exceptionally low miss rates of 5.01% (KNN) in the Cervical Cancer dataset, 4.49% (GBT) in the Lung Cancer dataset, and 3.10% (KNN) in the Dermatology dataset.

References

Alhassan, A. M., & Wan Zainon, W. M. N. (2021). Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. IEEE Access, 9, 87310-87317. https://doi.org/10.1109/ACCESS.2021.3083840

Ayesha, S., Hanif, M. K., & Talib, R. (2021). Performance enhancement of predictive analytics for health informatics using dimensionality reduction techniques and fusion frameworks. IEEE Access, 10, 753-769. https://doi.org/10.1109/ACCESS.2021.3046633

Coello, C. A. C., & Castillo Tapia, M. G. (2021). Cultural algorithms for optimization. In Handbook of AI-based Metaheuristics (pp. 219-238). CRC Press.

Durairaj, M., & Poornappriya, T. S. (2020). Why feature selection in data mining is prominent? A survey. In Proceedings of the International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019 (pp. 123-134). Springer International Publishing. https://doi.org/10.1007/978-3-030-68295-7_11

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 20-30. https://doi.org/10.48275/jscdm.v2i1.119

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1), 20-30. https://doi.org/10.48275/jscdm.v2i1.119

Islam, M. T., & Xing, L. (2021). A data-driven dimensionality-reduction algorithm for the exploration of patterns in biomedical data. Nature Biomedical Engineering, 5(6), 624-635. https://doi.org/10.1038/s41551-021-00677-1

Kostick, K. M., et al. (2021). A principal components analysis of factors associated with successful implementation of an LVAD decision support tool. BMC Medical Informatics and Decision Making, 21, 1-14. https://doi.org/10.1186/s12911-021-01532-0

Meng, X. (2021). Optimization of cultural and creative product design based on simulated annealing algorithm. Complexity, 2021, Article 5538251. https://doi.org/10.1155/2021/5538251

Nanga, S., et al. (2021). Review of dimension reduction methods. Journal of Data Analysis and Information Processing, 9(3), 189-231. https://doi.org/10.4236/jdai.2021.93011

Patra, S. S., et al. (2021). Emerging healthcare problems in high-dimensional data and dimension reduction. In Advanced Prognostic Predictive Modelling in Healthcare Data Analytics (pp. 25-49). CRC Press. https://doi.org/10.1201/9780429275765-3

Poornappriya, T. S., & Durairaj, M. (2019). High relevancy low redundancy vague set based feature selection method for telecom dataset. Journal of Intelligent & Fuzzy Systems, 37(5), 6743-6760. https://doi.org/10.3233/JIFS-179152

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: A review. Artificial Intelligence Review, 54(5), 3473-3515. https://doi.org/10.1007/s10462-020-09853-y

Tripathy, B. K., Sundareswaran, A., & Ghela, S. (2021). Unsupervised learning approaches for dimensionality reduction and data visualization. CRC Press.

UCI Machine Learning Repository. (n.d.). Cervical cancer risk factors. Retrieved from https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

UCI Machine Learning Repository. (n.d.). Dermatology. Retrieved from https://archive.ics.uci.edu/dataset/33/dermatology

UCI Machine Learning Repository. (n.d.). Lung cancer. Retrieved from https://archive.ics.uci.edu/dataset/62/lung+cancer