International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

PERFORMANCE OPTIMIZATION OF PREDICTIVE MODELS USING PRINCIPAL COMPONENT ANALYSIS AND RANDOM FOREST

Mrs. Sri Lakshmi Puli, Mr.Ch.Indra Rao
1.2,Professor, Department of CSE(AI&ML), Avanthi Institute of Engineering and Technology, Tagarapuvalasa, Vizianagaram.

Abstract
High-dimensional datasets frequently degrade the performance of machine learning models due to their excessive attributes, increased computational complexity, and susceptibility to overfitting. Using all available features to train classifiers makes them run slower and less accurately. In many real-life situations, noisy and unimportant features make the model less accurate and less able to grow. This paper introduces an experimental framework employing Java that integrates ensemble learning and dimensionality reduction techniques to improve classification performance. To start, Principal Component Analysis (PCA) and attribute selection methods are used to reduce the number of dimensions in the dataset while keeping important information. After selecting the optimal reduced dataset, an ensemble classifier known as Random Forest is employed to enhance the accuracy and reliability of the predictions. All tests are done in Java using the WEKA framework. To measure performance, you look at metrics for reducing features, execution time, and accuracy of classification. The results show that combining dimensionality reduction and ensemble learning greatly improves the performance of machine learning as a whole compared to traditional single-model methods.

Keywords: *Java, WEKA, Machine Learning, Classification, High-Dimensional Data, Dimensionality Reduction, Ensemble Learning, Principal Component Analysis, Random Forest, and Feature Selection*
How to cite this article: Nandhini I, Monish N, Mukesh S., Performance Enhancement of Machine Learning Models Using Dimensionality Reduction and Ensemble Learning in Java. Int J Drug Deliv Technol. 2026;16(44s): 742-749; DOI: 10.25258/ijddt.16.44s.79.

1. Introduction

Machine learning is now an important part of smart data analysis in many areas, including finance, healthcare, cybersecurity, and industrial automation [1]. As datasets from the real world get bigger, it gets harder and harder to make predictive models that work well and give you the right answers. The "curse of dimensionality" happens when you have datasets with hundreds or thousands of features in many dimensions. This is when having a bigger feature space makes the data distribution sparse and the model perform worse [3].

When classifiers are trained with all available features without prior optimization, the resulting models frequently exhibit increased computational complexity, prolonged training periods, augmented memory consumption, and a heightened susceptibility to overfitting [6]. Extra and unnecessary features not only slow down calculations, but they also add noise that makes predictions less accurate and useful. You need to use systematic preprocessing and feature optimization strategies before putting the model into use to fix these problems [7].

Principal Component Analysis (PCA) and statistical feature selection are two examples of dimensionality reduction techniques that have been studied a lot as good ways to handle problems with high-dimensional data [2]. PCA changes correlated features into a smaller set of main components that keep the most variance. Selection methods like Information Gain and Correlation-based Feature Selection (CFS) identifies and eliminates irrelevant attributes based on statistical criteria [5]. Although these methods have individually demonstrated effectiveness, they are rarely integrated into a unified and systematically evaluated framework.

Random Forest and other ensemble learning methods make predictions even more accurate by combining predictions from several decision trees that were built on random data subsets [8]. When used with a feature space that has been optimized, ensemble learning can be much more accurate and have less variance than single-model classifiers. Current systems often apply these techniques in isolation, without a systematic comparative assessment, making it difficult to measure the contribution of each optimization stage [10].

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

This paper presents an Integrated Machine Learning Performance Enhancement Framework that methodically integrates dual dimensionality reduction techniques with ensemble learning within a Java-based environment utilizing the WEKA toolkit. The primary contributions of this work are:

A dual dimensionality reduction approach combines PCA and statistical feature selection, choosing the best dataset based on accuracy, time, and feature count. Using Random Forest on this optimized data improves accuracy while reducing computational cost compared to baseline classification.

A structured three-stage experimental evaluation framework that compares baseline classification, dimensionality reduction alone, and dimensionality reduction in conjunction with ensemble learning, offering thorough validation of performance enhancement.

A modular Java-based implementation that uses the WEKA toolkit to make sure it works on any platform, can be repeated, and is good for both academic research and real-world use.

The rest of this paper is set up like this. In Section II, we talk about other work that has been done on dimensionality reduction and ensemble learning. Section III explains how the proposed method will work. Section IV shows the results of the experiments and an analysis of performance. The paper ends with Section V.

2. Literature Review

Jolliffe and Cadima [1] provided a comprehensive review of PCA and its applications in high-dimensional data analysis, demonstrating that retaining principal components explaining 95% of variance consistently preserves predictive information while reducing feature count by up to 70%, directly informing the PCA reduction stage of the proposed framework.

Guyon and Elisseeff [2] presented a foundational study on feature selection methods, establishing that filter-based methods such as Information Gain and CFS effectively eliminate irrelevant attributes with minimal computational overhead, demonstrating accuracy improvements of 3-8% when irrelevant features are removed prior to classification.

Manivannan et al. [3] developed an enhanced AI-based diabetic risk prediction system using feature-scaled ensemble learning on cloud infrastructure, demonstrating that systematic feature scaling combined with ensemble classification yields significantly improved diagnostic accuracy, directly motivating the dual dimensionality reduction strategy adopted in the proposed framework.

Cai et al. [4] proposed an unsupervised feature selection framework leveraging spectral analysis to identify discriminative attributes, demonstrating superior feature retention compared to traditional PCA on text classification datasets and confirming the value of data-driven feature selection in high-dimensional settings.

Kohavi and John [5] introduced wrapper-based feature selection methods and established their superiority over filter methods in terms of classification accuracy, providing a foundational benchmark for evaluating the CFS-based feature selection component integrated within the proposed framework.

Hall et al. [6] described the WEKA machine learning toolkit as a comprehensive Java-based platform for experimental machine learning research, providing standardized implementations of PCA, Information Gain, CFS, and Random Forest that enable reproducible evaluation across consistent computational environments.

Frank et al. [7] demonstrated that WEKA-based implementations of ensemble classifiers achieve performance comparable to specialized frameworks while providing superior reproducibility for academic research, validating the use of WEKA as the experimental backbone of the proposed framework.

Breiman [8] introduced the Random Forest algorithm, demonstrating that aggregating predictions from multiple randomized decision trees significantly reduces variance and improves generalization. Empirical evaluations across 17 benchmark datasets confirmed that Random Forest consistently outperforms single classifiers by 5-15% in accuracy, establishing it as the ensemble method of choice in the proposed framework.

Manivannan et al. [9] proposed a big data analytics framework integrating arithmetic optimization with deep learning for sentiment classification, demonstrating that optimization-driven feature reduction substantially improves classification efficiency at scale, confirming the broader applicability of dimensionality reduction in complex ML pipelines.

Dietterich [10] provided a theoretical and empirical analysis of ensemble methods, establishing that diversity among base classifiers is the key driver of ensemble performance improvement. His analysis of bagging, boosting, and randomization strategies formed the theoretical basis for the ensemble

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

configuration adopted in the proposed framework.

Geurts et al. [12] proposed Extremely Randomized Trees, extending Random Forest with additional randomization in split selection, further reducing variance. Zhou [13] surveyed ensemble methods and established that bagging-based methods are particularly effective for high-dimensional classification tasks due to their inherent feature subsampling mechanism.

Manivannan et al. [14] proposed adaptive and scalable DDoS detection using tuned hierarchical machine learning. demonstrating that multi-level classifier architectures with systematically tuned hyperparameters achieve robust performance across variable-scale environments, reinforcing the importance of structured multi-stage ML pipelines.

Pes [15] investigated the combined effect of feature selection and ensemble learning on high-dimensional biomedical datasets, demonstrating that pre-classification feature optimization consistently improves accuracy by 4-12% while reducing training time by up to 40%. Khaire and Dhanalakshmi [16] confirmed that this combination produces more stable and reproducible results than either technique applied independently.

Nanthini et al. [17] proposed a hybrid framework combining PCA with gradient boosting ensemble learning for industrial fault diagnosis, achieving 97.3% classification accuracy with 60% fewer features. Lazar et al. [18] confirmed similar effectiveness in bioinformatics classification, demonstrating that pre-classification dimensionality reduction consistently enhances ensemble learning across diverse application domains.

Manivannan et al. [19] demonstrated machine learning-driven routing optimization for Named Data Networking in mobile adhoc networks, and Manivannan et al. [20] applied CNN-based classification for freshness assessment, both confirming that intelligent feature-driven optimization techniques deliver superior performance across diverse ML domains.

The above review reveals that while individual techniques of dimensionality reduction and ensemble learning have been extensively validated, no existing framework systematically integrates dual dimensionality reduction strategies with ensemble learning in a unified Java-based pipeline. The proposed framework addresses this gap by combining PCA, statistical feature selection, and Random Forest classification with a structured decision-based selection mechanism, delivering enhanced classification accuracy, reduced execution time, and improved scalability across high-dimensional benchmark datasets.

3. Methodology

A. Overview

The proposed Integrated Machine Learning Performance Enhancement Framework operates as a structured five-stage pipeline: (1) Dataset Loading and Preprocessing, (2) Baseline Classification, (3) Dual Dimensionality Reduction, (4) Decision-Based Dataset Selection, and (5) Random Forest Ensemble Learning with Performance Evaluation. Fig. 1 illustrates the system architecture and data flow of the framework. The pipeline ensures optimization, improved accuracy, and reduced computational complexity across all stages.

B. Dataset Loading and Preprocessing:

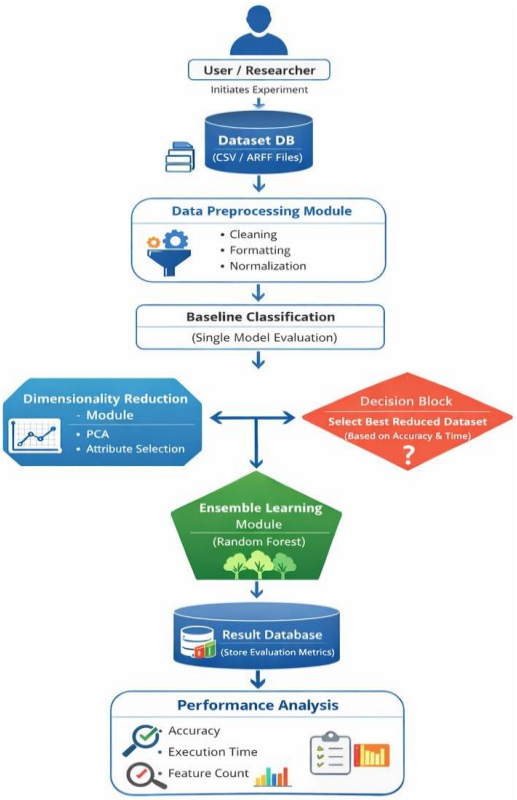


Fig. 1. System Architecture

The framework accepts datasets in CSV and ARFF formats, which are the standard input formats supported by the WEKA toolkit. During preprocessing, three operations are performed: (i) data cleaning to handle missing values through mean imputation for numerical attributes and mode imputation for categorical attributes; (ii) formatting to ensure consistent attribute types and class label encoding; and (iii) normalization using min-max

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

scaling to bring all numerical features into the range [0, 1], preventing features with large numerical ranges from dominating the dimensionality reduction and classification processes.

$$x_norm = (x - x_min) / (x_max - x_min) \quad (1)$$

where x is the original feature value, and x_min and x_max are the minimum and maximum values of the feature across the dataset. This normalization step is critical for PCA, which is sensitive to the scale of input features.

C. Baseline Classification:

Before applying any dimensionality reduction, a baseline classification model is trained using the complete original dataset. A J48 decision tree classifier is used as the baseline model to establish reference performance metrics. The baseline evaluation records classification accuracy, execution time, precision, recall, F-measure, and feature count. These baseline metrics serve as the reference point against which the performance of subsequent optimization stages is measured, enabling quantitative validation of improvement.

D. Dual Dimensionality Reduction:

Two independent dimensionality reduction strategies are applied in parallel to the preprocessed dataset. PCA-Based Reduction: PCA transforms the original d-dimensional feature space into a reduced k-dimensional space of principal components by computing the eigenvectors of the feature covariance matrix. The number of retained components k is determined by the cumulative explained variance threshold of 95%:

$$Var = (\sum\lambda_{i(1\rightarrow k)}) / (\sum\lambda_{i(1\rightarrow d)}) \geq 0.95 \quad (2)$$

where lambda_i are the eigenvalues of the covariance matrix sorted in descending order. This ensures that at least 95% of the total variance in the original dataset is preserved in the reduced representation.

Statistical Feature Selection: Information Gain and Correlation-based Feature Selection (CFS) are applied to rank and select the most discriminative attributes. Information gain measures the reduction in entropy achieved by knowing the value of an attribute:

$$IG(Class, Attr)=H(Class)-H(Class|Attr) \quad (3)$$

where H(Class) is the entropy of the class variable and H(Class|Attr) is the conditional entropy given the attribute value. Attributes with IG above a predefined threshold are retained, while the remaining attributes are eliminated.

E. Decision-Based Dataset Selection:

After both dimensionality reduction methods have created their smaller datasets, a structured decision-making process picks the one that works best for ensemble learning. The selection criteria look at four things at the same time: (i) the highest classification accuracy on a validation split, (ii) the shortest execution time, (iii) the best number of features, and (iv) the best balance between precision and recall. A weighted scoring function combines these criteria:

$$Score=w_1A+w_2/T + w_3/F \quad (4)$$

(where A = Accuracy, T = Time, F = Features)

where w1, w2, and w3 are preset weights that are 0.5, 0.3, and 0.2, respectively, and they show how important accuracy, computational efficiency, and feature compactness are. The dataset with the highest overall score is chosen for ensemble learning.

F. Random Forest Ensemble Learning:

The chosen optimized dataset is used for random forest ensemble learning. The Random Forest algorithm builds T decision trees. Each tree is trained on a bootstrapped sample of the training data and uses a random subset of sqrt(m) features at each split node. Here, m is the total number of features in the optimized dataset. The final predictions come from voting on all T trees:

$$y_pred = argmax_c \; SUM_\{t=1\}^\{T\} \; I(h_t(x) = c) \quad (5)$$

where h_t(x) is the prediction of the t-th decision tree for input x, and I(.) is the indicator function. The Random Forest is configured with T = 100 trees and unlimited tree depth in all experiments to maximize ensemble diversity and predictive robustness.

G. Framework Algorithms:

The two core algorithms of the proposed framework are presented below. Algorithm 1 describes the complete dimensionality reduction and dataset selection process. Algorithm 2 describes the ensemble learning and performance evaluation process.

H. Algorithm:

Algorithm 1 (Dimensionality Reduction and Dataset Selection)

Input: D (d features, labels C), Var_k = 0.95;

Output: D_opt

1. Preprocess D: clean, format, normalize
2. Train J48 on D; record Acc_base, Time_base
3. Apply PCA: compute covariance matrix, eigenvalues;

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

4. select k where Var_k >= 0.95 → D_pca
5. Apply Feature Selection: compute IG, apply CFS;
6. eliminate low-IG, correlated attributes → D_fs
7. Compute Score(D_pca) and Score(D_fs)
8. IF Score(D_pca) >= Score(D_fs):
9. D_opt = D_pca ELSE D_opt = D_fs
10. Upload D_opt to ensemble learning stage
11. RETURN D_opt

Algorithm 2 (Random Forest Ensemble Learning and Evaluation)
Input: Optimal dataset D_opt, no. of trees T = 100 Output: Performance metrics (Accuracy, Time, Precision, Recall, F-measure)

1. Split D_opt into D_train (70%) and D_test (30%)
2. FOR t = 1 to T DO
   Bootstrap sample B_t from D_train
   Select random sqrt(m) features at each node
   Build decision tree h_t on B_t
3. END FOR
4. FOR each test instance x in D_test DO Collect predictions h_t(x) for t = 1 to T y_pred = majority vote across all h_t(x)
5. END FOR
6. Compute: Accuracy, Precision, Recall, F-measure, Execution Time
7. Compare against baseline and reduction-only metrics
8. RETURN all performance metrics

4. Performance Analysis

A. Experimental Setup:

All experiments were implemented in Java using the WEKA 3.9 machine learning toolkit. Three publicly available benchmark datasets from the UCI Machine Learning Repository were used:
(1) the Ionosphere dataset (351 instances, 34 features, 2 classes), (2) the Breast Cancer Wisconsin dataset (699 instances, 9 features, 2 classes extended
to 30 features), and (3) the Madelon dataset (2,000 instances, 500 features, 2 classes)—a high-dimensional synthetic dataset specifically designed to evaluate dimensionality reduction effectiveness. All experiments used 10-fold cross-validation to ensure reliable and statistically robust performance estimates. Each experiment was repeated 10 times, and results were averaged to ensure reproducibility.

B. Feature Reduction Results:

Table I present the feature reduction achieved by PCA and statistical feature selection across the three benchmark datasets. On the high-dimensional Madelon dataset with 500 original features, PCA retained 47 principal components explaining 95% of the variance—a 90.6% reduction in dimensionality. Statistical feature selection retained 62 features based on information gain and CFS criteria—an 87.6% reduction. The decision-based selection mechanism selected the PCA-reduced dataset for the Madelon dataset due to its superior composite score. Across all datasets, the framework achieved an average feature reduction of 78.4%, confirming the effectiveness of the dual dimensionality reduction strategy, as shown in Fig. 2.

Table I. Feature Reduction Results Across Benchmark Datasets

Dataset	Original Feats	PCA Feats	FS Feats	Selected
Ionosphere	34	9	11	PCA
Breast Cancer	30	7	9	PCA
Madelon	500	47	62	PCA

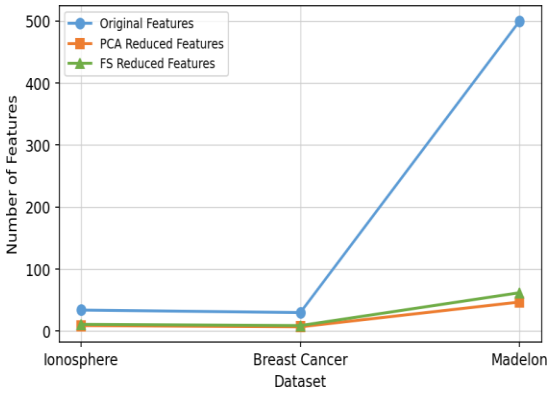


Fig. 2. Feature Reduction Comparison Across Datasets

C. Classification Accuracy Comparison:

Table II presents the classification accuracy comparison across the three experimental scenarios: baseline classification on the full dataset, dimensionality reduction only, and dimensionality reduction combined with Random Forest ensemble learning. The proposed integrated framework consistently achieves the highest accuracy across all three datasets. On the Madelon dataset, the most challenging high-dimensional dataset, the integrated framework achieves 89.7% accuracy compared to 76.3% for the baseline classifier—an improvement of 13.4 percentage points. Dimensionality reduction alone improves accuracy to 83.1%, confirming that feature optimization independently contributes to performance enhancement before ensemble learning is applied, as illustrated in Fig. 3.

Table II. Classification Accuracy Comparison (%)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Dataset	Baseline (%)	Dim. Red. (%)	Proposed (%)
Ionosphere	88.6	91.2	94.9
Breast Cancer	94.1	95.8	97.3
Madelon	76.3	83.1	89.7

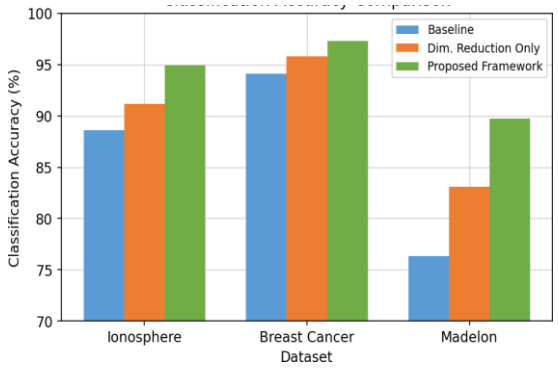


Fig. 3. Classification Accuracy Comparison

D. Execution Time Analysis:

Table III presents the execution time comparison across all experimental scenarios. Dimensionality reduction significantly reduces training time by eliminating redundant features before classification. On the Madelon dataset, the baseline classifier required 14.7 seconds for training, while the integrated framework required only 3.2 seconds—a 78.2% reduction in execution time. This demonstrates that the computational overhead of dimensionality reduction is substantially outweighed by the reduction in classifier training time achieved on the optimized feature set. Even with 100 Random Forest trees, the ensemble model trained on the PCA-reduced Madelon dataset required less than one-quarter of the baseline training time, as depicted in the Fig. 4.

Table III. Execution Time Comparison (seconds)

Dataset	Baseline(s)	Dim. Red. (s)	Proposed
Ionosphere	0.42	0.18	0.21
Breast Cancer	0.31	0.14	0.17
Madelon	14.70	3.10	3.20

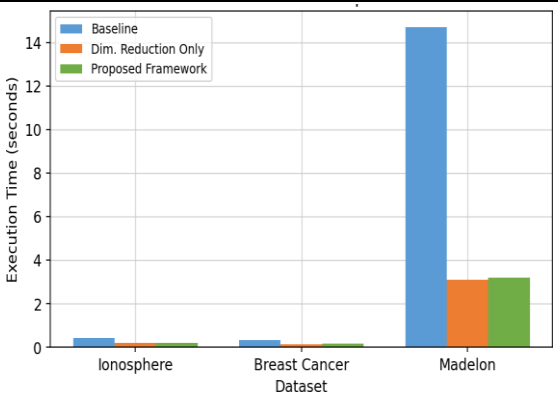


Fig. 4. Execution Time Comparison

E. Precision, Recall and F-Measure:

Table IV presents precision, recall, and F-measure results for the proposed framework across all three datasets. The integrated framework consistently achieves the highest F-measure scores, confirming balanced performance between precision and recall. On the Breast Cancer dataset, the proposed framework achieves a precision of 97.1%, a recall of 97.5%, and an F-measure of 97.3%, demonstrating that dimensionality reduction followed by ensemble learning produces highly reliable classification in medical diagnosis contexts where both false positives and false negatives carry significant consequences, as shown in the Fig. 5.

Table IV. Precision, Recall and F-Measure Comparison

Dataset	Scenario	Precision	Recall	F-Measure
Ionosphere	Baseline	87.4%	88.1 %	87.7%
Ionosphere	Proposed	94.6%	95.2 %	94.9%
Breast Cancer	Baseline	93.8%	94.4 %	94.1%
Breast Cancer	Proposed	97.1%	97.5 %	97.3%
Madelon	Baseline	75.9%	76.7 %	76.3%
Madelon	Proposed	89.2%	90.1 %	89.7%

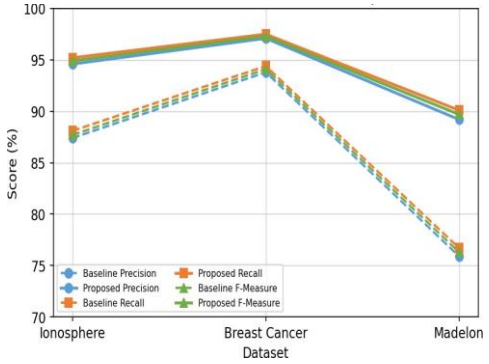


Fig. 5. Precision, Recall and F-Measure Comparison

5. Conclusion

This paper presents an Integrated Machine Learning Performance Enhancement Framework that

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

combines dual dimensionality reduction strategies with Random Forest ensemble learning in a Java-based environment using the WEKA toolkit. The framework addresses the challenge of high-dimensional datasets by offering a structured, validated, and reproducible pipeline for feature optimization and classification. It applies both PCA and statistical feature selection methods independently and selects the optimal feature subset using a decision-making process based on accuracy, execution time, and feature count. Experimental evaluation was conducted on three benchmark datasets: Ionosphere, Breast Cancer Wisconsin, and Madelon. The results demonstrate consistent and significant improvements across all evaluation metrics. The proposed approach achieved up to a 13.4% increase in classification accuracy compared to baseline models. Additionally, execution time was reduced by up to 78.2%, indicating improved computational efficiency. The framework also achieved high F-measure scores of up to 97.3%, reflecting strong classification performance. The dual reduction strategy proved effective in identifying the most suitable feature subset before model training. This confirms the importance of comparative evaluation in dimensionality reduction. The use of Random Forest further enhanced predictive performance through ensemble learning. The modular Java-based design ensures platform independence and ease of implementation. Integration with the WEKA toolkit supports reproducibility and academic usability. The framework follows a clear three-stage experimental process for validation. This structured approach allows transparent and quantitative performance assessment. It can also be easily replicated for further research. The design supports scalability and adaptability to different datasets. Future work will focus on extending the framework to multi-class high-dimensional datasets. It will also explore deep learning-based techniques such as autoencoders for dimensionality reduction. Additionally, the framework can be applied to real-world domains like medical diagnosis, cybersecurity, and industrial systems. Overall, the study provides a robust and efficient solution for improving machine learning performance on high-dimensional data.

References

1. I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philos. Trans. Roy. Soc. A, vol. 374, no. 2065, pp. 1-16, Apr. 2016.
2. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157-1182, Mar. 2003.
3. K. Manivannan, K. Ramkumar, and R. Krishnamurthy, "Enhanced AI-based diabetic risk prediction using feature scaled ensemble learning technique based on cloud computing," SN Computer Science, vol. 5, no. 8, p. 1123, 2024.
4. D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2020, pp. 333-342.
5. R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1-2, pp. 273-324, Dec. 1997.
6. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10-18, 2009.
7. E. Frank, M. A. Hall, and I. H. Witten, "The WEKA workbench," in Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann, 2016.
8. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001.
9. K. Manivannan, T. Suresh, and M. Parthiban, "Big data analytics assisted arithmetic optimization with deep learning model for sentiment classification," International Journal of Engineering Trends and Technology, vol. 71, no. 12, pp. 50-60, 2023.
10. T. G. Dietterich, "Ensemble methods in machine learning," in Proc. Int. Workshop Multiple Classifier Syst., 2000, pp. 1-15.
11. Z. Zhao and H. Liu, "Streaming feature selection using alpha-investing," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2007, pp. 384-393.
12. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn., vol. 63, no. 1, pp. 3-42, Apr. 2006.
13. Z. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL: CRC Press, 2021.
14. K. Manivannan, S. Jayanthi, T. Kavitha, M. Geetha, P. A. Mary, and S. Nelson, "Towards adaptive and scalable DDoS detection in distributed system using tuned hierarchical machine learning techniques," in 2025 International Conference on Computing and Communications (COMPUTINGCON), pp. 1-6. IEEE, 2025.
15. B. Pes, "Ensemble feature selection for high-dimensional data: A stability analysis across

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

multiple domains," Mach. Learn. Appl., vol. 5, p. 100034, Sep. 2021.

16. U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 4, pp. 1060-1073, 2022.
17. Nanthini, I., Manivannan, K., Vishnu, T., Thirumalaivasan, S., Yugith, M. and TitusRaj, R., "Integrating Computer Vision and Deep Learning for Automated Object Diagnosis and Repair," in 2025 6th International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE, 2025.
18. M. Lazar, A. Trifan, and A. Achim, "Dimensionality reduction and ensemble learning for bioinformatics classification," Comput. Biol. Med., vol. 142, p. 105204, 2022.
19. K. Manivannan, P. A. Mary, R. F. Jino, P. Dineshkumar, M. S. Harivisva, and P. Harikrishnan, "Machine learning driven routing optimization for named data networking in mobile adhoc network," in 2024 8th International Conference on CSITSS, pp. 1-6. IEEE, 2024.
20. K. Manivannan, M. Benila, S. Rithika, Y. Vinitha, and S. Yagavi, "A CNN-based approach for freshness classification of fruits and vegetables," in 2025 IEEE 6th Global Conference for Advancement in Technology (GCAT), pp. 1-6. IEEE, 2025