Project 2: Dimensionality Reduction

Ravi Teja Anumula (UFID: 73522006)
Dept. of Electrical and Computer Engineering
University of Florida
Gainesville, USA

***Abstract*—This project focuses on applying dimensionality reduction and machine learning techniques to classify satellite imagery containing "ship" and "no-ship" labels. Utilizing a dataset of RGB images, we aim to develop and evaluate machine learning models for efficient and accurate ship detection. The project encompasses various stages, including data visualization, feature reduction, model training, and performance evaluation. Initial steps involve training classifiers without dimensionality reduction to establish baseline performance metrics, followed by implementing dimensionality reduction through Principal Component Analysis (PCA) and manifold learning algorithms. These techniques enhance model interpretability and computational efficiency by reducing the feature space.**

We explore and compare classifiers on both the original and reduced feature spaces, analyzing performance and inference time to identify the optimal dimensionality reduction and classification pipeline. Visualization tools, such as 2D embeddings of the reduced feature space, provide insights into the clustering and separability of image classes. A final comprehensive evaluation across unseen full-scene satellite images tests the model's ability to locate ships accurately within complex backgrounds, thus demonstrating the effectiveness of automated ship detection in real-world applications.

***Index Terms*—Dimensionality Reduction, Principal Component Analysis (PCA), Manifold Learning, Hyperparameter tuning**

I. Introduction

This project aims to enhance the accuracy and efficiency of automated ship detection in satellite images, supporting applications like coastal security, logistics, and environmental monitoring. Using machine learning and dimensionality reduction, we developed a model to classify images as "ship" or "no ship." The diverse dataset of labeled satellite images allows for a reliable detection system to aid data-driven maritime operations.

Our main objectives include:

A. Enhanced Detection via Predictive Modeling

B. Dimensionality Reduction for Computational Efficiency

C. Improved Classification through Model Optimization

D. Real-World Application in Maritime Monitoring

II. Methodology

This methodology outlines the steps from data preparation to model testing. Below is a summary of the main components:

A. Data Preparation

The dataset contains 4,000 labeled satellite images (80x80 RGB) of "ship" or "no-ship." Due to high dimensionality, dimensionality reduction techniques were explored to enhance computational efficiency.

B. Model Training with Full Feature Set

To establish a baseline, three classifiers were trained on the full dataset without dimensionality reduction

a) Random Forest Classifier: An ensemble method suited for high-dimensional data.

b) Logistic Regression: A linear model for binary classification

c) Decision Tree Classifier: A simple model that creates interpretable rules for classification. Each model underwent hyperparameter tuning to optimize performance, serving as a control for comparison with reduced-feature models.

C. Dimensionality Reduction with PCA

PCA was applied to retain 90% of data variance, compressing the feature space and potentially reducing noise. The classifiers were retrained with PCA-transformed data, followed by hyperparameter tuning to assess the impact of dimensionality reduction on performance

D. Dimensionality Reduction with Manifold Learning

ISOMAP, a nonlinear technique, was also applied to capture the data's intrinsic structure. Models were trained and tuned on ISOMAP-transformed data to evaluate performance in a lower-dimensional, non-linear space.

E. Hyperparameter Tuning and Cross-Validation

GridSearchCV and cross-validation were used for tuning each model-dimension reduction pair to avoid overfitting and achieve optimal performance.

F. Performance Evaluation and Comparison

Models were assessed on accuracy, F1 score, and training time. Comparison metrics included confusion matrices and classification reports to highlight strengths and weaknesses, with misclassified samples examined for patterns.

G. Real-World Application Testing

The top-performing model was further tested on full satellite scenes to evaluate its generalization capabilities. This test simulated real-world conditions, allowing for an assessment of the model's effectiveness in detecting ships across varied and complex oceanic backgrounds.

III. Observations

A. Performance metrics without dimensionality reduction

Parameters	***Random Forest***	***Logistic Regression***	***Decision Tree***
Training Time	14.12 ms	9.25 ms	31.47 ms
Accuracy	1.0	1.0	1.0
F1_score	1.0	1.0	1.0
Precision	C0-1.0, C1-1.0	C0-1.0, C1-1.0	C0-1.0, C1-1.0
Recall	C0-1.0, C1-1.0	C0-1.0, C1-1.0	C0-1.0, C1-1.0

TABLE I
PERFORMANCE METRICS WITHOUT DIMENSIONALITY REDUCTION

Using the complete dataset without any dimensionality reduction (i.e., no PCA or ISOMAP), we can assess the performance of three classifiers: Random Forest, Logistic Regression, and Decision Tree.

Based on the table, the performance of the Random Forest, Logistic Regression, and Decision Tree models on the full feature dataset (without dimensionality reduction) demonstrates perfect classification results with 100 % accuracy, F1-score, precision, and recall across both classes (C0: ”No Ship” and C1: ”Ship”).

In summary, while all models perform exceptionally well on the training set, Logistic Regression stands out in terms of efficiency. However, further validation on test data is needed to ensure that these models do not suffer from overfitting.

B. Dimensionality Reduction with PCA

On performing the dimensionality reduction with PCA and observing the outcomes, it can be seen that we need atleast ’104’ principal components to get a cumulative explained ratio greater than 90% .

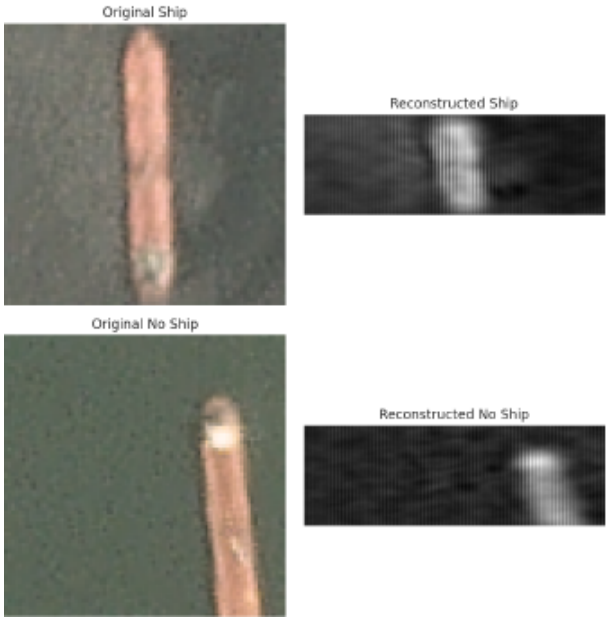

Fig. 2. Visualize examples of ”ship” and ”no ship” image reconstructions.

After applying the dimensionality reduction with a variance of over 90%, the images are reconstructed to visualize the differences prior and post reduction. Visualizations of reconstructed images for ”ship” and ”no-ship” classes using 104 components show that while the major features of the images are preserved, finer details might be lost due to dimensionality reduction.

The average RMSE (Root Mean Squared Error) of reconstruction was calculated for different component counts. As the number of PCA components increases, the RMSE decreases, indicating better reconstruction quality. For 104 components (90% variance), the RMSE was approximately 13.24, showing a balance between reduced dimensionality and reconstruction fidelity. A graph plotting RMSE against the number of components demonstrates this trend.

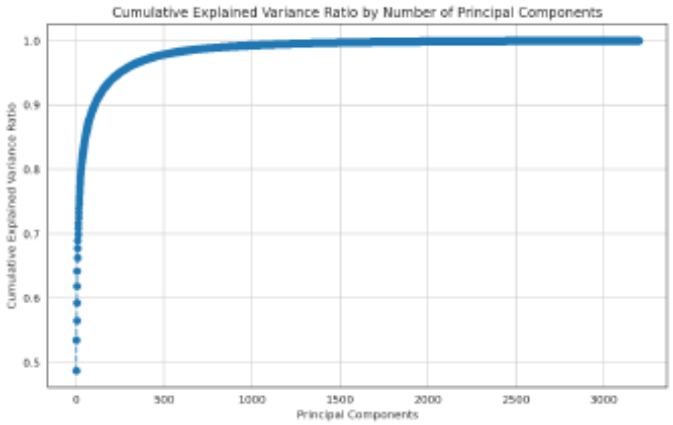

Fig. 1. Explained variance ratio by number of principal components.

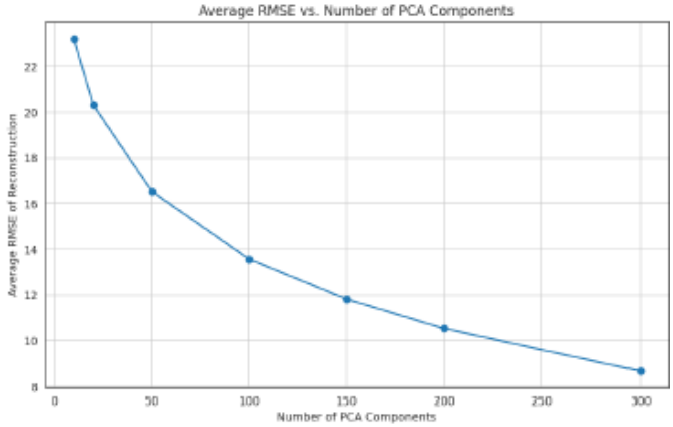

Fig. 3. Average RMSE vs Number of PCA Components

C. Performance metrics with dimensionality reduction via PCA

Parameters	***Random Forest***	***Logistic Regression***	***Decision Tree***
Training Time	7.11 ms	5.21 ms	4.49 ms
Accuracy	1.0	0.937	0.99
F1_score	1.0	0.87	0.989
Precision	C0-1.0, C1-1.0	C0-0.95, C1-0.89	C0-0.99, C1-1.0
Recall	C0-1.0, C1-1.0	C0-0.97, C1-0.85	C0-1.0, C1-0.98

TABLE II
PERFORMANCE METRICS WITH DIMENSIONALITY REDUCTION VIA PCA

On comparing the performance metrics in the table after applying dimensionality reduction via PCA across Random Forest, Logistic Regression, and Decision Tree models:

a) Random Forest: Maintains perfect accuracy, F1 score, precision, and recall across both classes, making it the best performer in both accuracy and robustness after PCA, though it has a slightly longer training time

b) Decision Tree: Shows near-perfect metrics, achieving fast training and strong performance, making it a close second to Random Forest in both efficiency and effectiveness.

c) Logistic Regression: While it benefits from a reduced training time, it experiences a noticeable drop in accuracy, F1 score, and recall for the "ship" class, making it less suitable for high-accuracy requirements in this classification task.

D. Performance metrics with dimensionality reduction via ISOMAP

a) Random Forest: The best performer, with perfect accuracy, F1 score, precision, and recall for both classes. Despite a slightly higher training time, its overall performance is unmatched.

b) Decision Tree: A close second, with high accuracy, F1 score, precision, and recall, particularly excelling in balance and reliability.

c) Logistic Regression: While it benefits from a shorter training time, its accuracy, F1 score, and recall for the "ship" class are lower, making it less reliable than the other models after PCA.

Parameters	***Random Forest***	***Logistic Regression***	***Decision Tree***
Training Time	13.42 ms	12.16 ms	12.41 ms
Accuracy	1.0	0.93	0.965
F1_score	1.0	0.86	0.927
Precision	C0-1.0, C1-1.0	C0-0.94, C1-0.90	C0-0.96, C1-0.97
Recall	C0-1.0, C1-1.0	C0-0.97, C1-0.82	C0-0.99, C1-0.89

TABLE III
PERFORMANCE METRICS WITH DIMENSIONALITY REDUCTION VIA ISOMAP

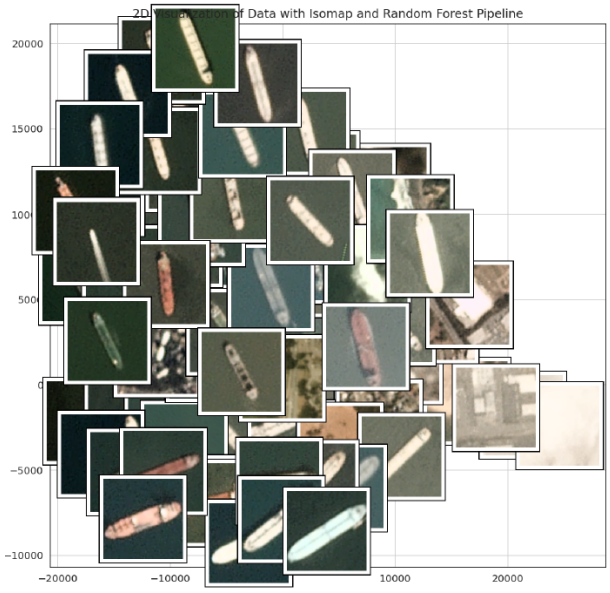

Fig. 4. 2D Visualization of Data with ISOMAP

During the dimensionality reduction with ISOMAP, the first two dimensions likely capture variations in the shape and orientation of the ships. Images of ships that have similar shapes and orientations are grouped close together in this 2D space. For example, ships that appear side-by-side with similar lengths and widths may cluster together, while images with ships at different angles may spread out along one axis. Some clusters of images show ships with different colors or textures. For instance, ships in varying shades (white, red, gray) or with different surroundings (water vs. land background) might be grouped based on color patterns or contrast in texture, indicating that these dimensions also capture color/texture information to some extent.

E. Confusion Matrices and misclassified samples

Models	***TP***	***FP***	***FN***	***TN***
Random Forest without Reduction	2400	0	0	800
Logistic Regression without Reduction	2400	0	0	800
Decision Tree without Reduction	2400	0	0	800
Random Forest with PCA	2400	0	0	800
Logistic Regression with PCA	2318	82	120	680
Decision Tree with PCA	2397	3	16	784
Random Forest with ISOMAP	2400	0	0	800
Logistic Regression with ISOMAP	2325	75	144	656
Decision Tree with ISOMAP	2376	24	88	712

TABLE IV
CONFUSION MATRICES FOR TRAINING DATA

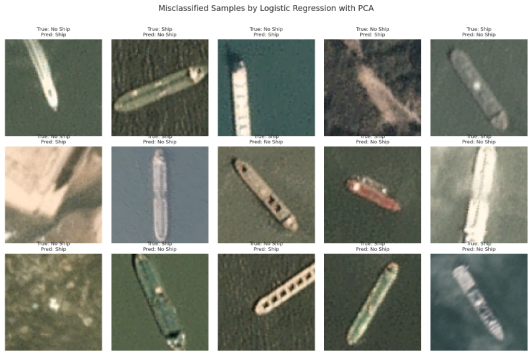

Fig. 5. Misclassified samples by Logistic Regression of PCA

a) Patterns in mis-classification: Some images of ships were misclassified as "No Ship" (e.g., row 1, column 1; row 1, column 4; row 2, column 2). These images often have ships that blend into the background due to similar colors or lower contrast, making them harder to distinguish. Similarly, some "No Ship" images that contain structures or water patterns resembling parts of a ship were classified as "Ship" (e.g., row 1, column 2; row 3, column 1). This shows the model may be influenced by shapes and colors that are typically associated with ships but appear in non-ship images as well.

b) Possible Solutions: Using a non-linear dimensionality reduction technique (like Isomap) might capture more complex, non-linear patterns that can help separate ships from non-ships, especially when backgrounds are similar. Techniques such as rotation, brightness adjustment, and contrast enhancement could help the model generalize better and become less sensitive to background noise and lighting differences. Adjusting hyperparamteter tuning could enhance the results as well.

F. Test performance

Models	***TP***	***FP***	***FN***	***TN***
Random Forest without Reduction	546	6	25	175
Logistic Regression without Reduction	555	45	33	167
Decision Tree without Reduction	563	37	32	168
Random Forest with PCA	599	1	28	172
Logistic Regression with PCA	571	29	39	161
Decision Tree with PCA	558	42	31	169
Random Forest with ISOMAP	577	23	37	163
Logistic Regression with ISOMAP	577	23	46	154
Decision Tree with ISOMAP	570	30	55	145

TABLE V
CONFUSION MATRICES FOR TEST DATA

Models	***Accuracy***	***F1 score***	***Precision***	***Recall***
Random Forest without Reduction	0.96	0.92	C0-0.96, C1-0.97	C0-0.99, C1-0.88
Random Forest with PCA	0.96	0.92	C0-0.96, C1-0.99	C0-1, C1-0.86
Random Forest with ISOMAP	0.925	0.844	C0-0.94, C1-0.88	C0-0.96, C1-0.81
Logistic Regression without Reduction	0.90	0.81	C0-0.94, C1-0.79	C0-0.93, C1-0.83
Logistic Regression with PCA	0.915	0.825	C0-0.94, C1-0.85	C0-0.89, C1-0.91
Logistic Regression with ISOMAP	0.91	0.82	C0-0.93, C1-0.87	C0-0.96, C1-0.77
Decision Tree without Reduction	0.91	0.83	C0-0.95, C1-0.82	C0-0.94, C1-0.84
Decision Tree with PCA	0.91	0.82	C0-0.95, C1-0.80	C0-0.93, C1-0.84
Decision Tree with ISOMAP	0.89	0.83	C0-0.91, C1-0.83	C0-0.95, C1-0.72

TABLE VI
PERFORMANCE METRICS FOR TEST DATA

Based on performance metrics and training/testing times, the random forest classifier shows overfitting, while logistic regression and decision tree classifiers demonstrate better generalization with 90% training accuracy. Logistic regression with PCA has low training time without compromising performance, making "model2_lr" (Logistic Regression with ISOMAP) the optimal choice.

G. Real-World Application Testing

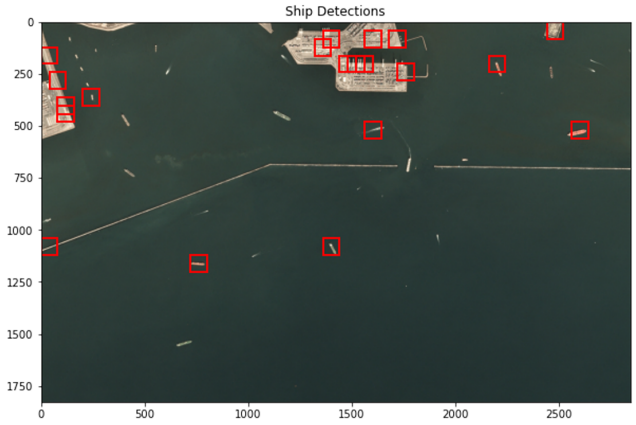

Fig. 6. Real world application testing

The output image shows the model's capability to detect and localize ships in a complex maritime scene, with red boxes marking areas classified as "ships." The model accurately identifies ships near docks and in open water, demonstrating its potential for real-world applications like maritime traffic monitoring and coastal security. The clear red boxes facilitate quick visualization, enhancing the efficiency of manual monitoring.