SYMBIOSIS INSTITUTE OF TECHNOLOGY (SIT)
Constituent of Symbiosis International (Deemed University), Pune
(Established under Section 3 of the UGC Act of 1956 vide notification number F-9-12/2001-U-3 of the Government of india)
Re-Accredited by NAAC with 'A++' Grade
॥वसुधैव कुटुम्बकम्॥

Project Report UML LAB

ON

EuroSAT Image Classification with Dimensionality Reduction Techniques and their Comparison

Under the Guidance of

Dr. Anupkumar Bongale

EuroSAT Image Classification with Dimensionality Reduction

By

Krisha shah 24070126512,

Janhavi doijad 23070126153

Table of Contents

1. Introduction

With the rapid increase in Earth observation systems and satellite imagery, classifying land use and land cover from remote sensing data has become essential for environmental monitoring, agricultural analysis, and urban planning. However, the high dimensionality of image data poses challenges in training efficient and generalizable models.

The EuroSAT dataset, derived from Sentinel-2 satellite data, provides a diverse collection of labeled land cover images. This project explores the use of dimensionality reduction techniques and transfer learning (ResNet50) to enhance the performance of traditional machine learning classifiers on EuroSAT data.

2. Objective

The core objectives of this project are to:

- Leverage **ResNet50**, a pretrained deep CNN, for extracting meaningful features from satellite images.
- Apply and compare multiple **dimensionality reduction** techniques — PCA, LDA, SVD, t-SNE, and MDS — to manage high-dimensional data efficiently.
- Train and evaluate a variety of **machine learning classifiers**, including Random Forest, SVM, XGBoost, and MLP, on the reduced feature space.
- Analyze the **trade-offs** between classification accuracy, computation time, and model complexity.

3. EuroSAT Dataset Description(https://github.com/phelber/EuroSAT)

The **EuroSAT dataset** is a publicly available satellite image dataset built from **Sentinel-2** satellite data provided by the European Space Agency (ESA). It is specifically curated for **land use and land cover classification** tasks using machine learning and deep learning techniques.

- **Total Images**: ~27,000
- **Image Size**: 64 × 64 pixels
- **Color Channels**: RGB (Red, Green, Blue)

- **Number of Classes**: 10 land use categories
- **Source**: Sentinel-2 satellite (multispectral imaging)
- **Resolution**: 10 meters per pixel

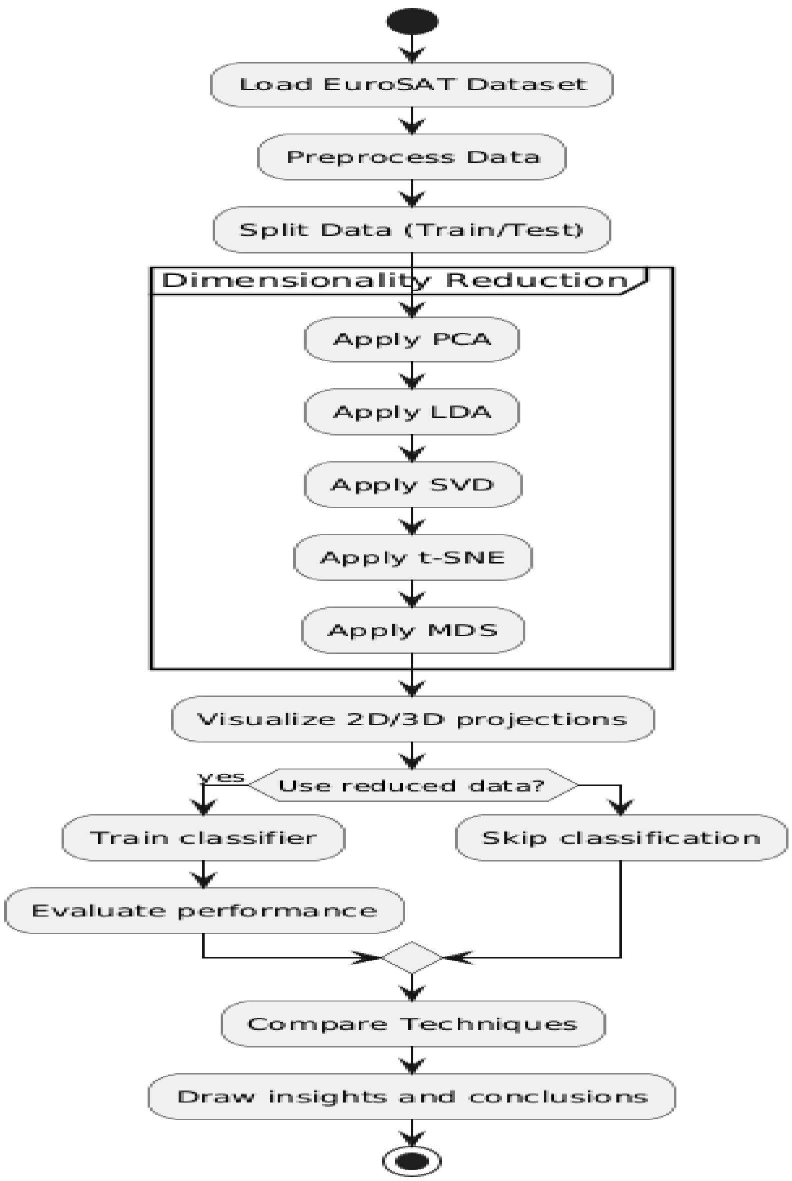
🏷️ **Land Use Classes:**

1. Annual Crop
2. Forest
3. Herbaceous Vegetation
4. Highway
5. Industrial
6. Pasture
7. Permanent Crop
8. Residential
9. River
10. Sea/Lake

4. Methodology

The project follows the pipeline below:

1. **Load the EuroSAT dataset** and preprocess all images (resize, normalize).
2. **Extract deep features using ResNet50** pretrained on ImageNet (transfer learning).
3. **Apply dimensionality reduction** on extracted features (PCA, LDA, etc.).
4. **Train classifiers** (Random Forest, SVM, XGBoost, MLP) using the reduced features.
5. **Evaluate model performance** using macro-averaged metrics and confusion matrices.
6. **Compare techniques** based on accuracy, interpretability, and computational cost.



5. Algorithms and Tech Stack

Tools and Libraries:

- Python
- Scikit-learn: ML models, metrics, dimensionality reduction
- XGBoost: Gradient boosting classifier
- TensorFlow / Keras: For loading pretrained ResNet50
- OpenCV: Image resizing and manipulation
- Matplotlib, Seaborn: Visualization

Why ResNet50?

Using a pretrained ResNet50 model allows us to extract semantically rich, spatially-informed features from satellite images without training a CNN from scratch. These features significantly improve classification performance when used with traditional ML classifiers.

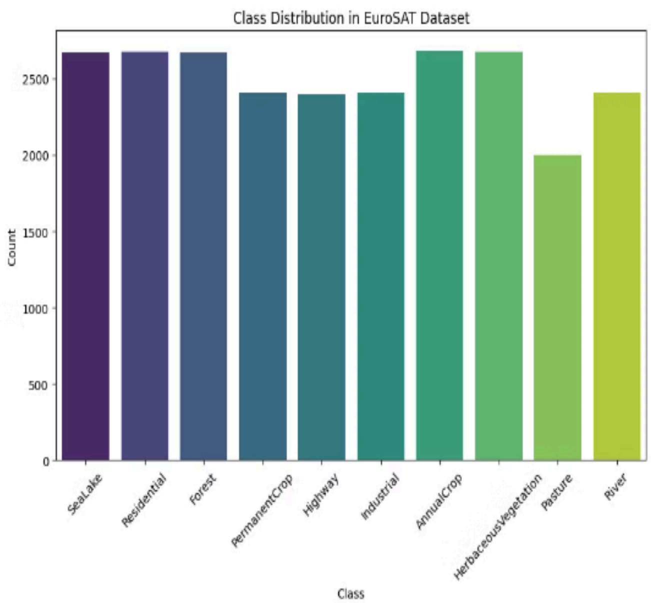
Dimensionality Reduction:

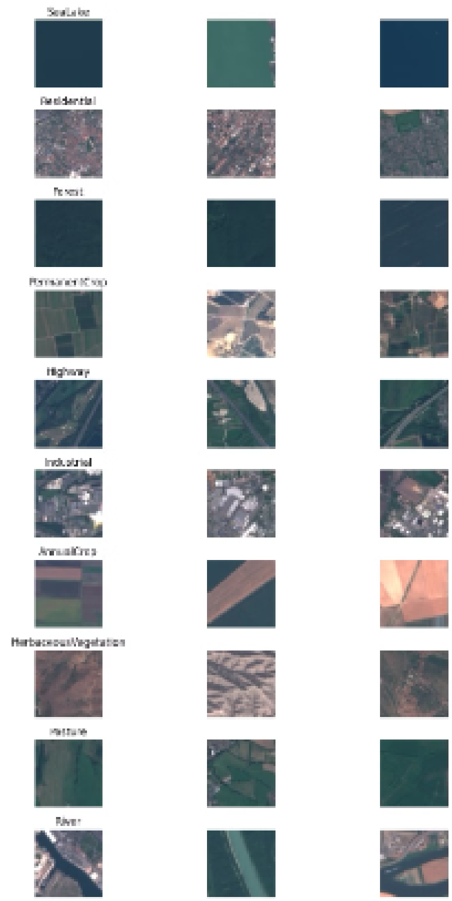
- PCA: Captures max variance, reduces redundancy
- LDA: Maximizes class separation (supervised)
- SVD: Extracts principal signal components
- t-SNE & MDS: Preserve structure for 2D visualization (not ideal for classification

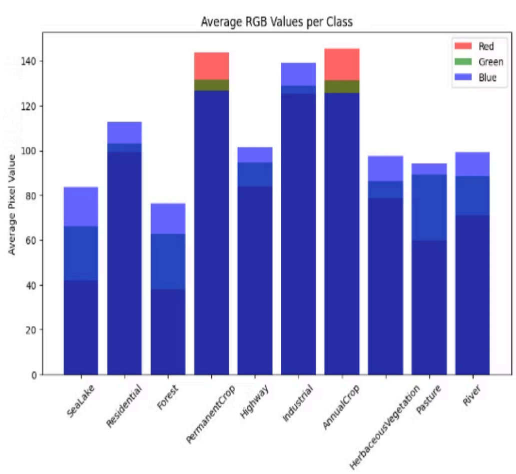
6. Dataset Exploration & Preprocessing

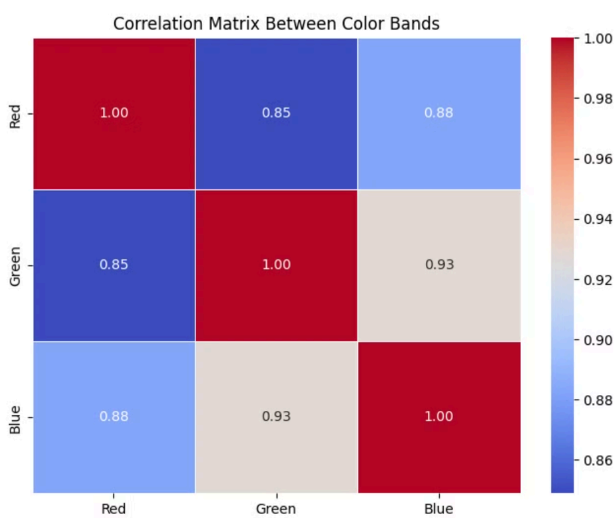
Dataset exploration includes verifying structure, inspecting sample images, checking image dimensions, analyzing class distribution, normalizing pixel values, analyzing color patterns, checking pixel intensity histograms, and creating a correlation matrix.

- Verified that all 10 classes have a roughly equal number of samples (~2,700)
- Displayed sample images to visually distinguish class characteristics.
- Analyzed RGB histograms and pixel distributions.
- Observed high feature correlation in raw image data, supporting the need for dimensionality reduction.


Class Distribution in EuroSAT Dataset
Count
0
500
1000
1500
2000
2500
SeaLake
Residential
Forest
PermanentCrop
Highway
Industrial
AnnualCrop
HerbaceousVegetation
Pasture
River
Class




Average RGB Values per Class
Red
Green
Blue
Average Pixel Value
0
20
40
60
80
100
120
140
SeaLake
Residential
Forest
PermanentCrop
Highway
Industrial
AnnualCrop
HerbaceousVegetation
Pasture
River


Correlation Matrix Between Color Bands
Red
Green
Blue
1.00
0.85
0.88
0.85
1.00
0.93
0.88
0.93
1.00
Red
Green
Blue
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86

7. Data Handling and Preprocessing

- **Image Resize**: All images resized to 64×64×3 to match ResNet50 input.
- **Normalization**: Applied Keras's preprocess_input() function to scale pixel values.
- **Feature Extraction**: Images passed through **ResNet50 (include_top=False, pooling='avg')** to obtain 2048-dimensional feature vectors.
- **Flattening & Reshaping**: Extracted vectors reshaped into (samples × 2048) format.
- **Label Encoding**: Converted categorical labels to numerical values.
- **Train-Test Split**: 80% for training, 20% for testing.

8. Implementation

After deep features were extracted from ResNet50, we applied dimensionality reduction using:

- **PCA**, **LDA**, **SVD** for classification
- **t-SNE**, **MDS** for data visualization

On these reduced features, we trained the following classifiers:

- **Random Forest**: Ensemble decision tree method, robust and interpretable.
- **SVM (RBF)**: Effective in high-dimensional spaces.
- **XGBoost**: Efficient gradient-boosted decision trees.
- **MLP**: Multi-layer Perceptron for learning nonlinear patterns.

Each model was trained using cross-validation and evaluated using standard classification metrics.

9. Dimensionality Reduction Techniques

- **PCA (Principal Component Analysis)**
  PCA reduces the dimensionality of a dataset by projecting it onto principal components that capture the maximum variance. By identifying the most significant features, it helps remove redundancy while retaining essential patterns in the data. This technique is particularly useful for handling large datasets, as it speeds up model training without significantly impacting performance.

- **LDA (Linear Discriminant Analysis)**
  LDA is a supervised method that enhances class separability by projecting data onto a new axis where classes are best distinguished. Unlike PCA, which focuses on variance, LDA optimizes the distinction between categories by maximizing the distance between class means while minimizing the variance within each class. This makes it highly effective for classification problems where well-defined target labels exist.
- **SVD (Singular Value Decomposition)**
  SVD decomposes a dataset into three matrices (U, Σ, V), revealing dominant patterns in the data. This technique is widely used in image processing and noise reduction, as it helps compress large datasets while preserving critical information. In applications like EuroSAT, SVD can efficiently extract key spatial features from satellite images.
- **t-SNE (T-Distributed Stochastic Neighbor Embedding)**
  t-SNE is primarily used for visualization, as it converts high-dimensional data into a 2D or 3D representation while preserving local structure. Unlike PCA, which emphasizes global variance, t-SNE focuses on maintaining relationships between similar data points. While highly effective for exploring complex data distributions, it is computationally expensive and not ideal for direct classification tasks.
- **MDS (Multidimensional Scaling)**
  MDS aims to preserve the original pairwise distances between data points while representing them in a lower-dimensional space. This technique is beneficial for analyzing similarities in spatial datasets, such as satellite images, where maintaining relationships between different geographical regions is crucial. It helps in visualizing patterns that may not be immediately apparent in high-dimensional space.

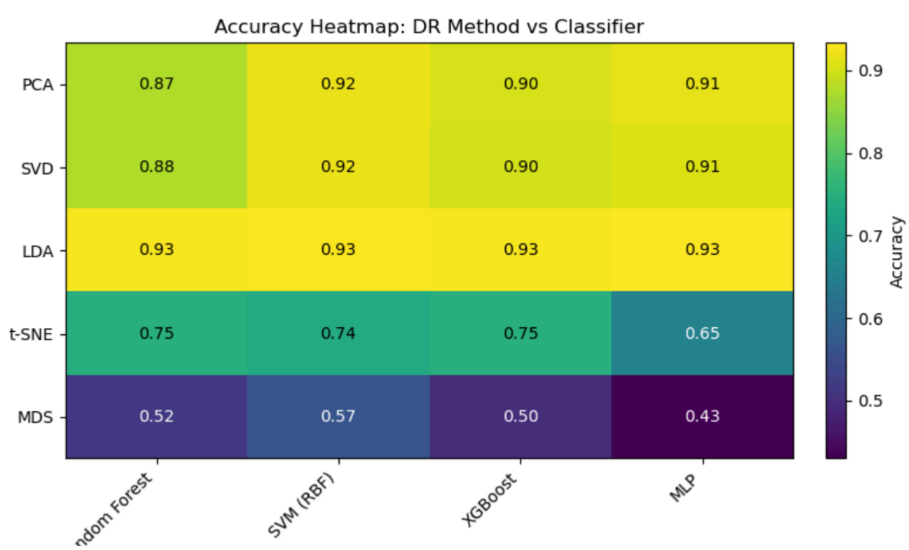
10. Feature Extraction with Transfer Learning

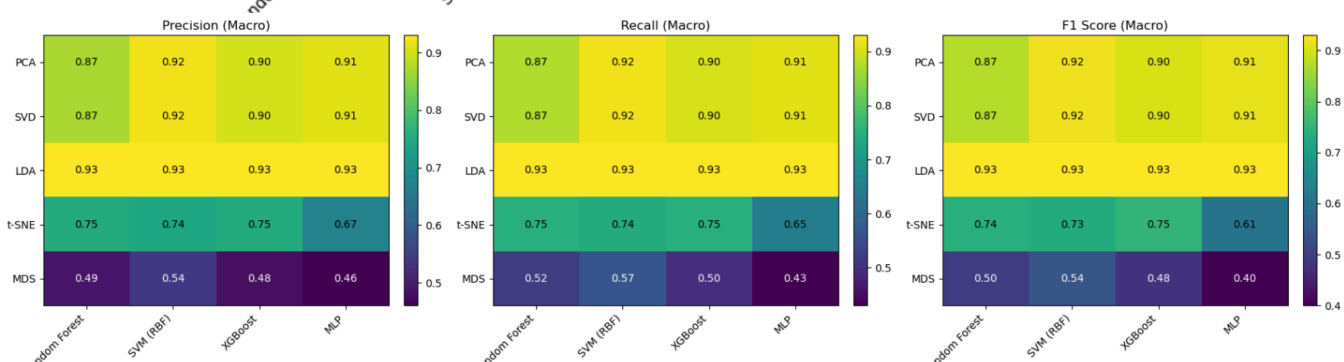
A core strength of our project lies in integrating **transfer learning** with the classification pipeline.

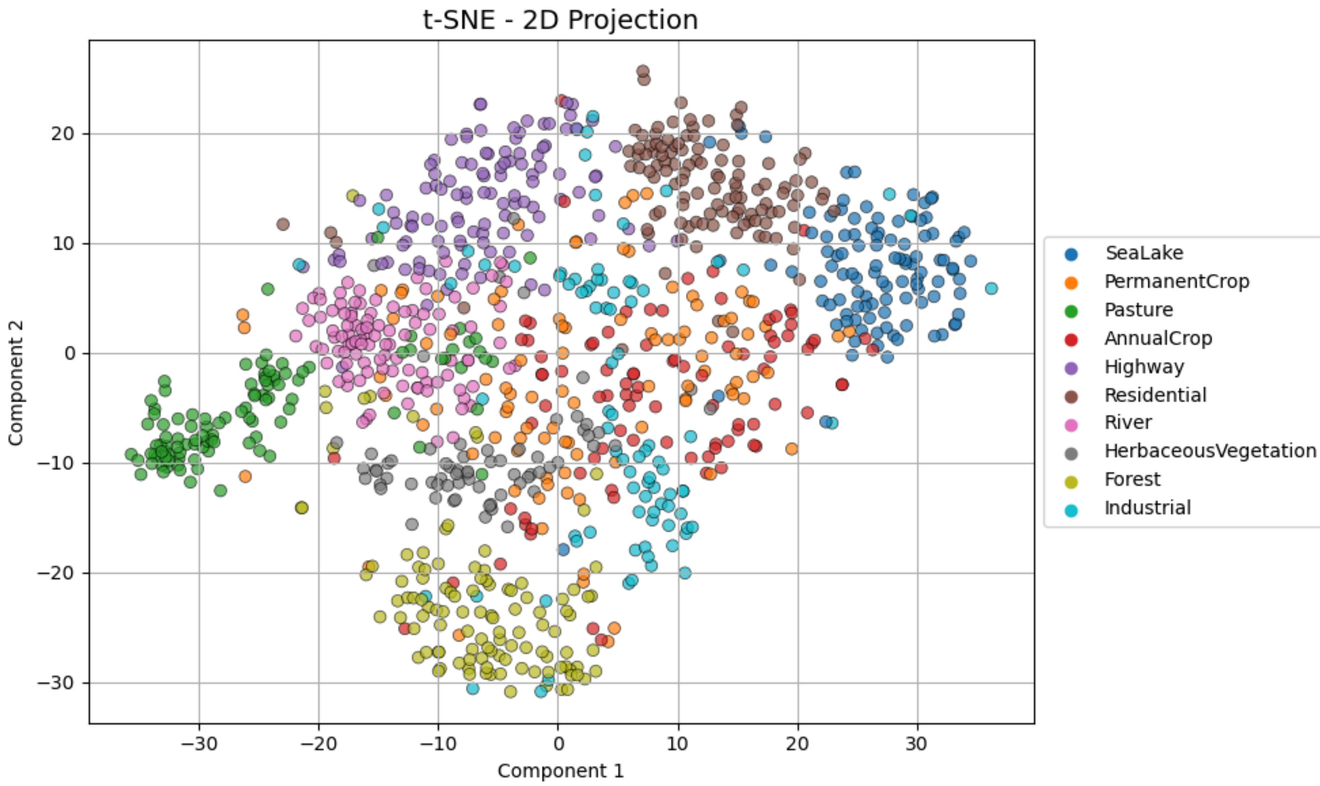
- We utilized **ResNet50 pretrained on ImageNet** to extract high-level features from each satellite image.
- Instead of training a full CNN, we used the **output of the global average pooling layer** (2048D vector) as a robust embedding for each image.
- This dramatically reduced noise and preserved meaningful spatial features, boosting the quality of input data for ML classifiers.

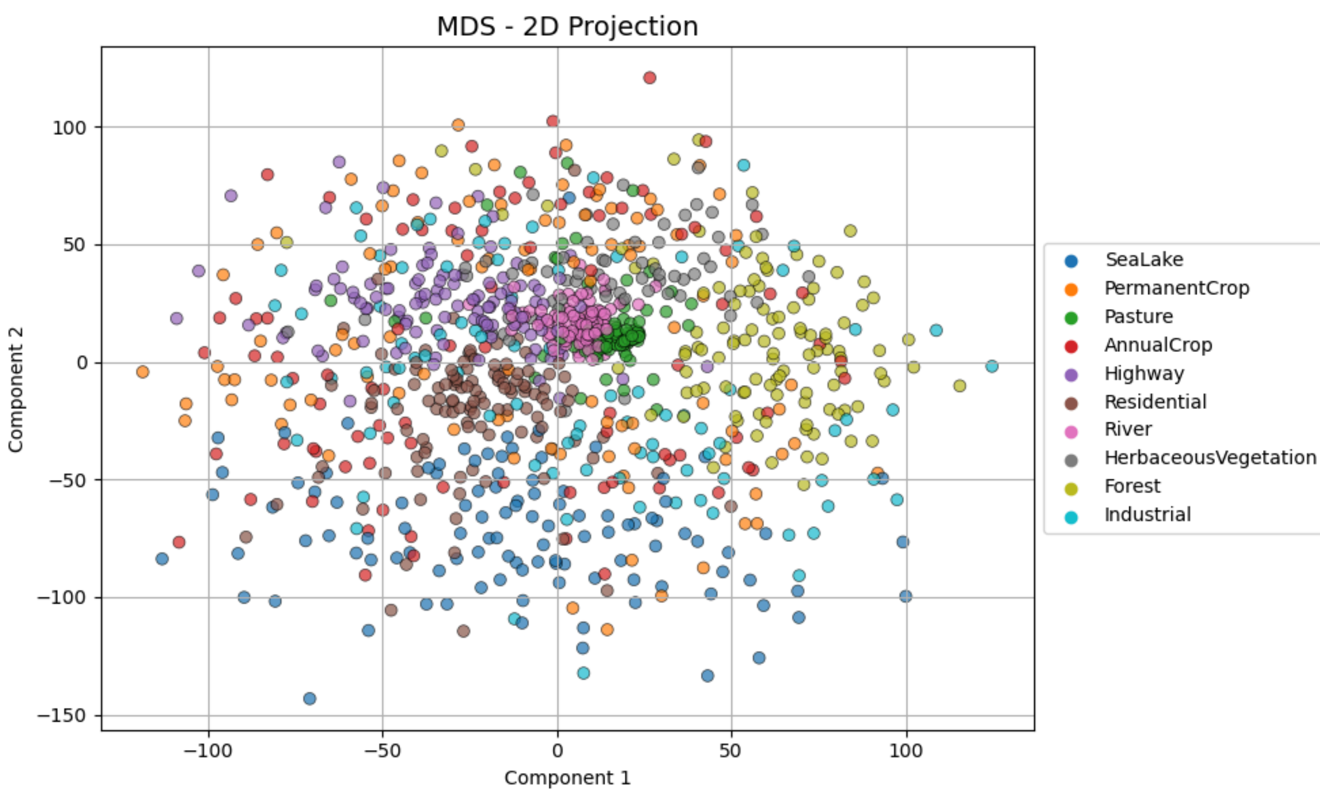
11. Evaluation and Results

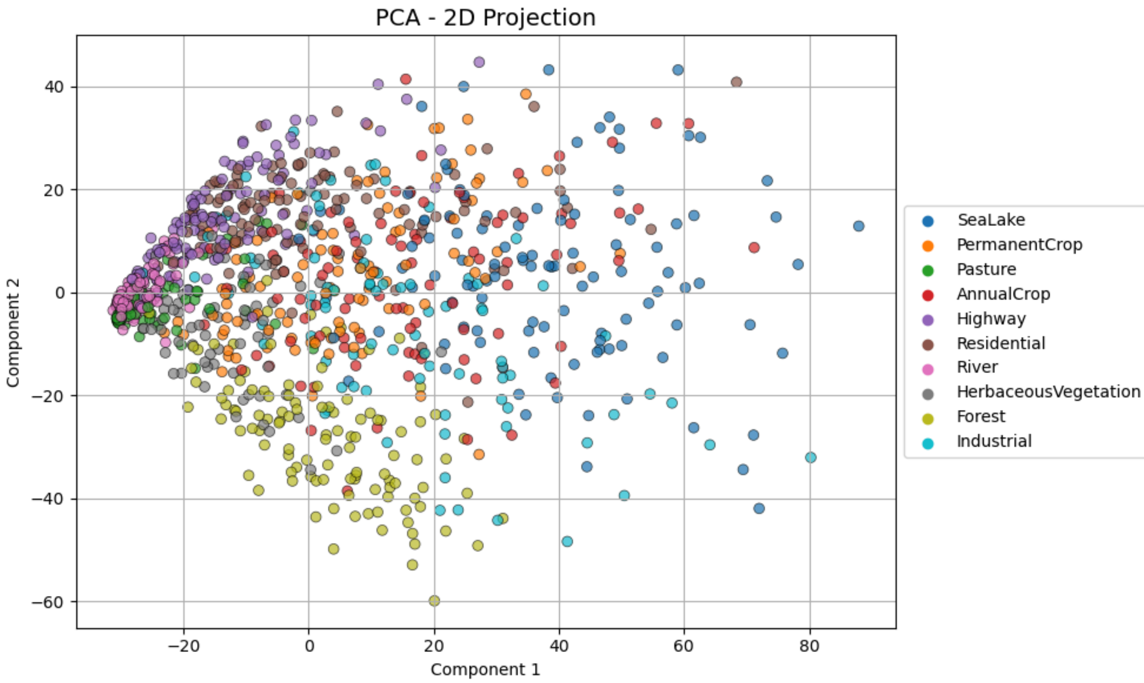
Performance is evaluated on the test set using various metrics, including accuracy, precision, recall, and F1-score. A confusion matrix is visualized to analyze class-specific performance, identifying potential areas for improvement. Comparing models based on these metrics provides insights into the effectiveness of different dimensionality reduction and classification techniques. Analyzing these results helps optimize the workflow for real-world applications.

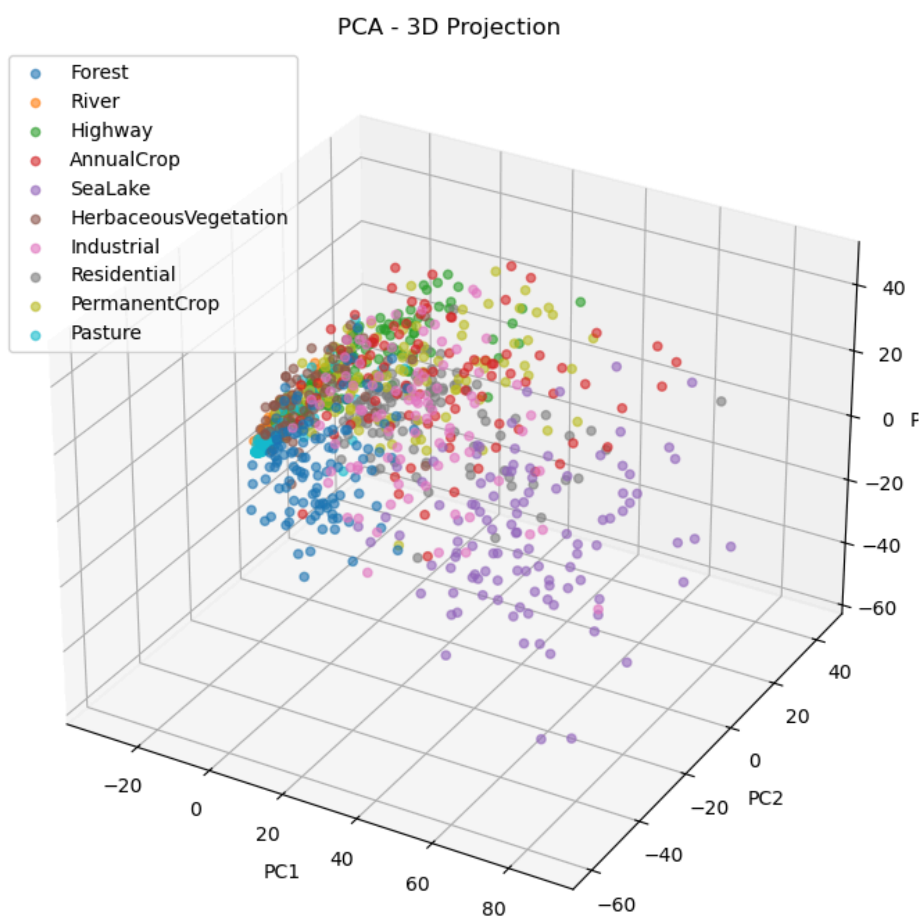


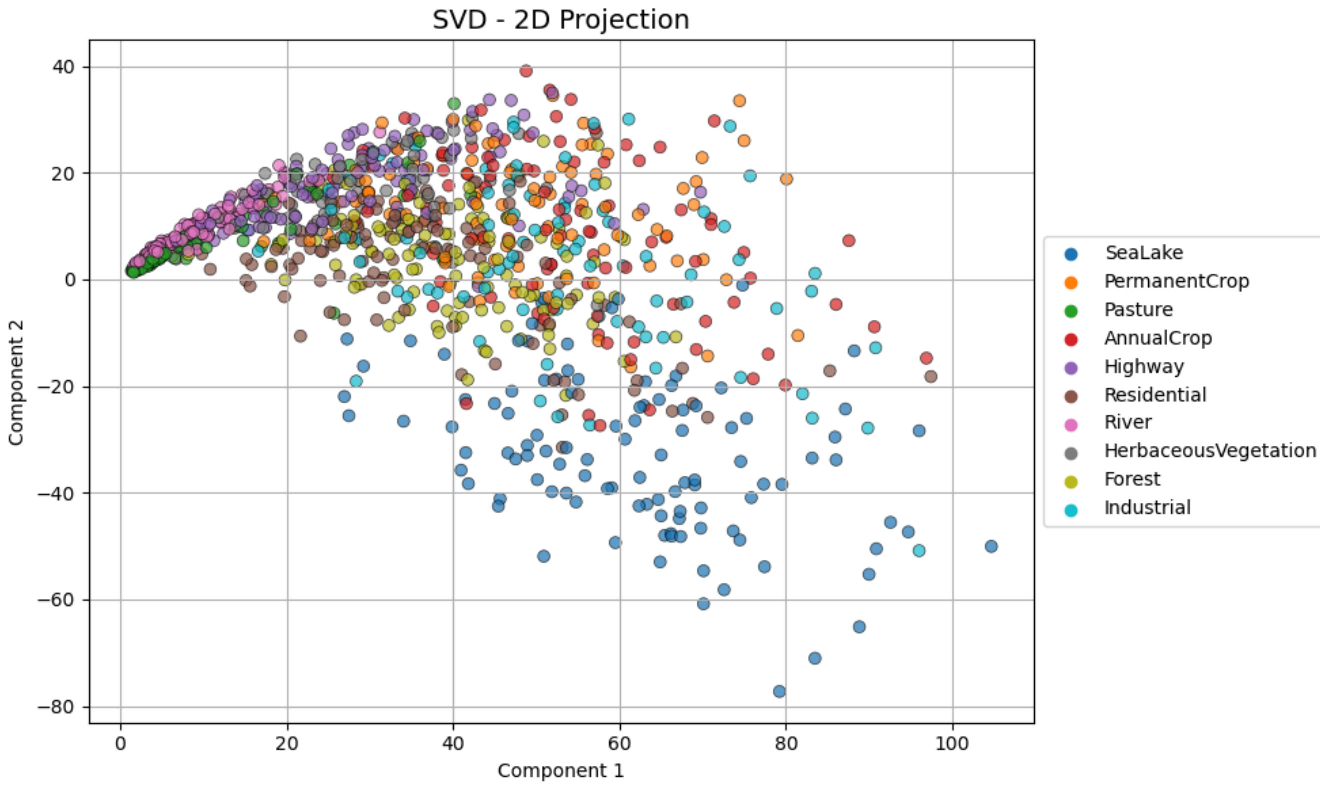



t-SNE - 2D Projection
Component 2
Component 1
SeaLake
PermanentCrop
Pasture
AnnualCrop
Highway
Residential
River
HerbaceousVegetation
Forest
Industrial

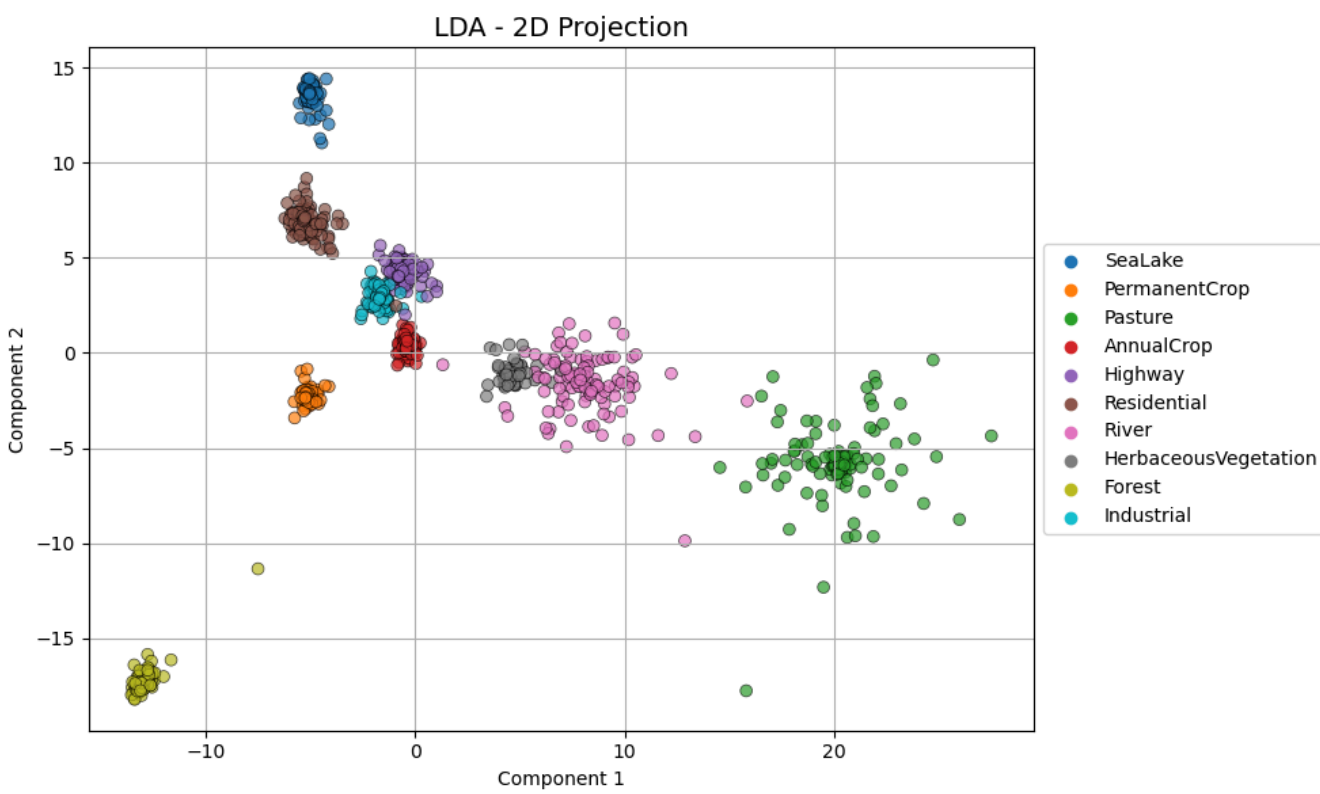

MDS - 2D Projection
Component 2
Component 1
SeaLake
PermanentCrop
Pasture
AnnualCrop
Highway
Residential
River
HerbaceousVegetation
Forest
Industrial


PCA - 2D Projection
Component 1
Component 2
SeaLake
PermanentCrop
Pasture
AnnualCrop
Highway
Residential
River
HerbaceousVegetation
Forest
Industrial


PCA - 3D Projection
Forest
River
Highway
AnnualCrop
SeaLake
HerbaceousVegetation
Industrial
Residential
PermanentCrop
Pasture
PC1
PC2



SVD - 2D Projection
Component 1
Component 2
SeaLake
PermanentCrop
Pasture
AnnualCrop
Highway
Residential
River
HerbaceousVegetation
Forest
Industrial

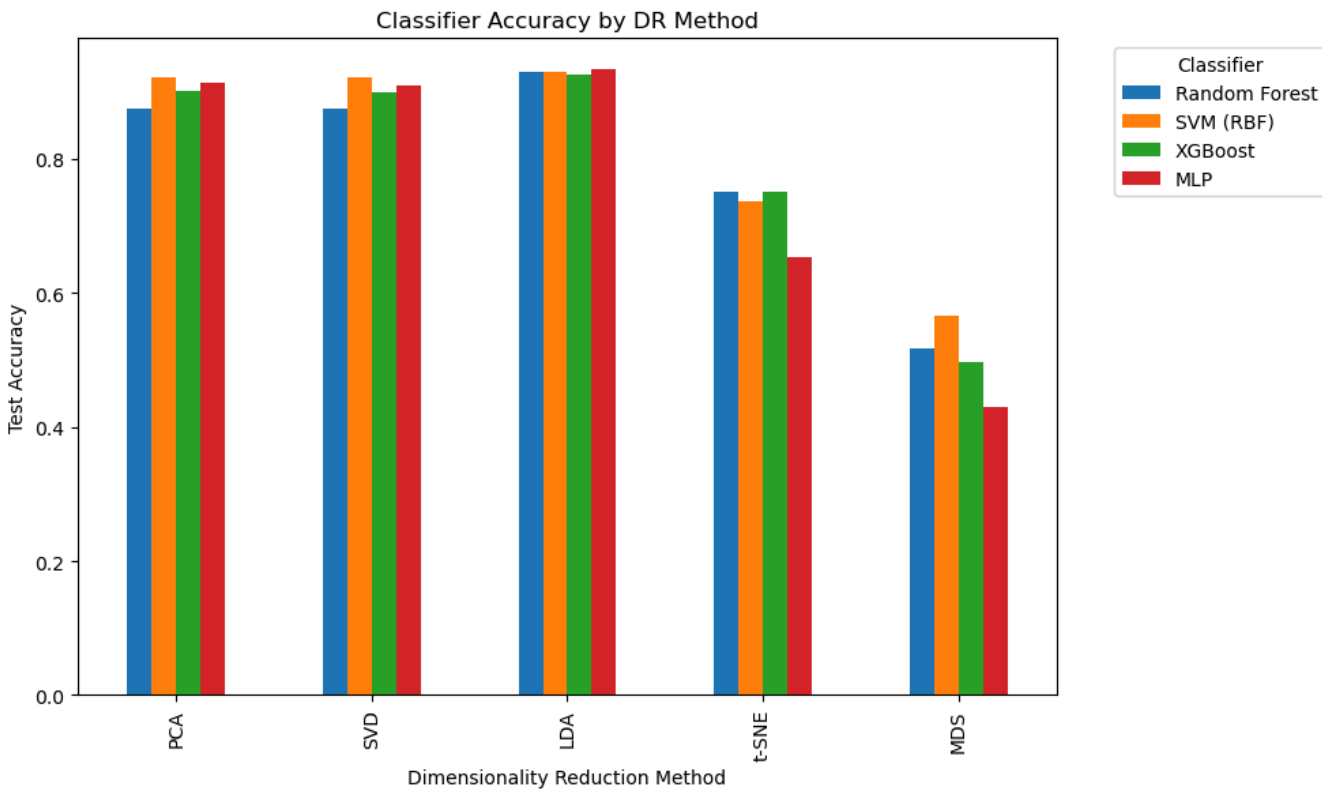


LDA - 2D Projection
Component 1
Component 2
SeaLake
PermanentCrop
Pasture
AnnualCrop
Highway
Residential
River
HerbaceousVegetation
Forest
Industrial

12. Conclusion

This project demonstrates that combining **deep feature extraction via ResNet50** with **dimensionality reduction techniques** results in a highly effective and scalable solution for satellite image classification.

- Transfer learning helped distill relevant features without requiring complex CNN training.
- Dimensionality reduction, especially **LDA**, improved classification performance and computational efficiency.
- The pipeline generalizes well and can be adapted for other remote sensing tasks or datasets.



13. Future Scope

- Use **Autoencoders** or **tuned CNNs** as nonlinear dimensionality reduction methods.
- Combine deep learning with ensemble learning for hybrid classification.
- Expand to **multispectral bands** beyond RGB for finer granularity.
- Deploy the model in real-time for edge-based satellite monitoring.

12. References

- Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). *EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification*. IEEE JSTARS.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.
- van der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE*. JMLR.
- Pedregosa et al., (2011). *Scikit-learn: Machine Learning in Python*. JMLR.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. CVPR.