COMPARISON OF PRINCIPAL COMPONENT ANALYSIS AND DIFFERENT BAND SELECTION METHODS FOR CLASSIFICATION OF CONSTRUCTION WASTE WITH HYPERSPECTRAL IMAGES

Lennard Wunsch[a, *], *Gunther Notni*[a,b]

[a]Ilmenau University of Technology, Ilmenau, Germany
[b]Fraunhofer Institute for Applied Optics and Precision Engineering IOF Jena, Jena, Germany
*Corresponding author: lennard.wunsch@tu-ilmenau.de

Abstract This work presents a machine learning pipeline for construction waste sorting utilizing spectral imaging and comparing different dimensionality reduction methods. Aiming to correctly classify over 90% of objects in our dataset, we applied band selection methods based on Mutual Information, Fisher's Score, Sequential Forward Selection, and Sequential Backward Selection. In addition, we examined Principal Component Analysis (PCA) and Categorical Maximum Spectral Difference. The performance of each pipeline is evaluated using metrics such as, accuracy, precision, recall, F1 Score and AUC-ROC

***Keywords*:** machine learning, pre-processing, spectral imaging, dimensionality reduction

1. INTRODUCTION

While RGB imaging is commonly used, its limited spectral resolution can hinder performance in tasks like construction waste sorting. Hyperspectral imaging overcomes this limitation by capturing rich spectral information and thus enabling precise material identification. However, its high dimensionality introduces the Hughes phenomenon, leading to overfitting and reduced model generalization. This study compares dimensionality reduction methods [1-4], Principal Component Analysis (PCA) [5] and various band selection techniques, to address the Hughes phenomenon and to optimize classification performance in construction waste sorting using a dataset of 26,266 hyperspectral images. Evaluation includes accuracy, precision, recall, F1 Score, and AUC-ROC.

2. DIMENSIONALITY REDUCTION METHODS

Dimensionality reduction can be achieved by either transforming the data into a lower-dimensional space or by removing redundant and irrelevant features. PCA is a transformation-based method that projects the data into a new coordinate system formed by principal components [5]. While PCA itself does not require dimensionality reduction, it enables it by ranking these components based on their variance. By selecting components with the highest variance and therefore information density the dimensionality of the data can be reduced while reserving as much of the important information as possible.

Feature Selection techniques do not change the data in any way but use metrics to rank all features. These techniques are categorized into filter-based, wrapper-based and embedded methods [1]. Filter-based methods evaluate all features independently of any learning algorithm and use statistical approaches instead.

Wrapper-based methods are based on learning algorithms to evaluate the performance of different subsets of features. The subsets are ranked based on the performance of the learning algorithm. Compared to filter-based methods this technique is computationally more expensive. Embedded methods perform feature selection as an integral part of the model training process. The model inherently selects or emphasizes the most relevant features while learning, often through mechanisms like regularization or feature importance scoring.

While Mutual Information (MI) [2], Fisher's Score (FS) [3] and categorical maximum spectral difference are filter-based methods, Sequential Backward Selection (SBS) and Sequential Forward Selection (SFS) are wrapper-based methods [4]. MI [2] measures the mutual dependence between features and a target. Wavelengths with high mutual information with the target label are more informative for distinguishing between classes and thus can be considered more meaningful for classification. FS [3] is a statistical approach that evaluates each feature individually based on its ability to discriminate between classes. It utilizes the ratio of between-class variance and within-class variance.

Categorical Maximum Spectral Difference describes an empirical method which calculates the spectral channels with the maximum difference between mean spectral for all classes to a base spectrum.

SFS and SBS find the optimal subset of bands by comparing the performance of different machine learning models, which are trained on these subsets. They differ in their starting point. SFS begins with an empty set of features and adds them one by one, while SBS starts with all features and removes them one by one. [4]

3. METHODOLOGY

After using each of these methods to select a subset of three wavelengths of the first three components in case of PCA, the hyperspectral data were reduced to only three wavelengths. The original data were captured using a HySpex SWIR-384 hyperspectral imaging system. The images contain 288 spectral channels from 930 nm to 2500 nm. 26,266 objects were captured divided into 5 different classes, resulting in a multi-class problem. The dataset is highly imbalanced, with the largest class containing 10,372 instances and the smallest class having only 1,535. During pre-processing image pixel values were normalized to a range of 0 to 1 to improve model performance. Additionally, all images were resized to 224 x 224 pixels to match the input size expected by the model architecture.

Afterwards the dimensionality reduction methods were applied to the data resulting in six distinct datasets. Since

many of these methods evaluate dependencies and correlations adjacent spectral bands were grouped together and the centre band was used. The resulting datasets were used for training, validation and testing of the same machine learning architecture. A VGG19-based convolutional neural network [6] was used as a backbone with a custom fully connected neural network head. The dataset was split into 80% training, 10% validation and 10% testing. The VGG19 base was initialized using pre-trained weights from ImageNet dataset, while the custom neural network head was randomly initialized.

All layers of the model were set to be trainable, allowing fine-tuning of both the base and classification head. Training was performed over 50 epochs using a batch size of 32 and a learning rate of 1e-6, optimized with Adam. Sparse categorical cross-entropy was used as the loss function, appropriate for multi-class classification with integer labels. Model performance was evaluated using standard metrics: accuracy, macro-precision, macro-recall, F1 score, and AUC-ROC [7]. Accuracy offers an overall measure of effectiveness, while macro-precision and macro-recall assess the model's ability to correctly identify and retrieve positive instances across all classes, irrespective of class frequency. AUC-ROC was computed using a One-vs-Rest strategy, averaging the area under the ROC curve for each class against all others.

4. RESULTS

This section presents the findings of the study based on the methods and analysis described previously. While MI, FS and Categorical Maximum Spectral Difference resulted in different spectral channels, SBS and SFS resulted in the same spectral regions. Therefore, for evaluation purposes these two methods were combined.

Table 1: Selected spectral bands

Method	1st Channel	2nd Channel	3rd Channel
Spectral diff.	1686 nm	1953 nm	2200 nm
MI	986 nm	1002 nm	1035 nm
FS	1198 nm	1652 nm	1237 nm
SBS/SFS	1046 nm	1810 nm	1908 nm
PCA	All wavelengths considered		

Table 1 presents the selected channel for each dimensionality reduction method. While most methods avoid the water absorption peaks in NIR, SBS/SFS include the peak at around 1900 nm as third selected band.

Table 2: Overview of the evaluation results

Method	Accuracy [%]	Precision [%]	Recall [%]	F1 Score [%]	AUC-ROC [%]
Spectral diff.	95.59	99.54	99.58	99.56	97.31
MI	94.48	99.59	99.96	99.77	97.73
FS	94.75	99.67	99.88	99.77	98.14
SBS/SFS	95.93	99.63	99.88	99.75	97.91
PCA	91.93	99.63	99.42	99.52	97.68

Table 2 present all methods used with their respective metrics. With overall high accuracies exceeding 90% all models demonstrate robust performance aligned with the study's objectives. Categorical Maximum Spectral Difference and SBS/SFS achieve the highest accuracy. While the F1 Score and AUC-ROC of Categorical Maximum Spectral Difference is lower than those of MI and FS, SBS/SFS outperforms it across all metrics. However, SBS/SFS records the second-highest F1 Score, not surpassing the top-performing methods in that category. MI and FS both attain the best F1 Score, with FS achieving the highest AUC-ROC, and MI the highest recall. PCA records show the lowest performance across all metrics.

5. CONCLUSIONS

A machine learning pipeline for construction waste sorting is presented, including a comparison of different dimensionality reduction methods. With high metrics all presented methods are viable to solve the classification of construction materials. All band selection methods demonstrated superior performance compared to the transformation-based approach. This suggests that targeted band selection is more effective than relying on a broad acquisition of spectral data followed by subsequent transformations. These findings highlight the importance of selecting informative spectral bands to optimize model accuracy and efficiency.

Future research could explore additional transformation-based dimensionality reduction methods such as Independent Component Analysis (ICA) or Non-negative Matrix Factorization (NMF), as well as learning-based approaches, including autoencoder architectures.

ACKNOWLEDGMENTS

Thanks to Dr. Galina Polte for useful discussion about machine learning and feature selection methods.

This research was funded by BMBF/Projektträger Jülich, the grant numbers 03RU1U151(A, C, J),

03RU1U152(A, C, J) and 03RU1U153C of projects of the alliance RUBIN-AMI.

REFERENCES

[1] B. Venkatesh, J. Anuradha, **A Review of Feature Selection and its Methods**, Cybernetics and Information Technologies19(1), 2019, 3–26.

[2] P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, **Normalized Mutual Information Feature Selection**, IEEE Transactions on Neural Networks 20(2), 2009, pp. 189-201.

[3] Q. Gu, Z. Li, J. Han, **Generalized Fisher Score for Feature Selection**, CoRR, 2012, abs/1202.3725.

[4] R. Panthonga, A. Srivihok, **Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm**, Procedia Computer Science 72, 2015, pp. 162-169.

[5] M. Ringnér, **What is principal component analysis?** Nat Biotechnol 26, 2008, pp. 303–304.

[6] K. Simonyan, A. Zisserman, **Very Deep Convolutional Networks for Large-Scale Image Recognition**, 3rd International Conference on Learning Representations (ICLR 2015), 2015, 1-14.

[7] I. Goodfellow, Y. Bengio, A. Courville, **Deep Learning**, The MIT Press Montreal, 2016, ISBN: 978-0-262-03561-3