ESTIMATING IMPORTANCE OF HIGHLY CORRELATED FEATURES USING MATRIX DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperspectral images contain a large volume of source data that exhibits high correlations along neighboring spectral bands. This makes it necessary to select the most informative features among correlated groups of features to effectively solve various machine learning problems. A method of feature importance evaluation for hyperspectral image data is proposed. This method combines iterative training of decision tree classifiers based on spectral features with matrix decomposition to overcome sparsity. Decision trees provide intrinsic feature selection mechanism but only a small number of features are usually taken into account by the CART algorithm for training a single decision tree classifier instance. Furthermore when features are highly correlated (e.g., Pearson $\rho > 0.8$), tree-based methods like Random Forest or XGBoost arbitrarily assign importance to one feature while suppressing others, as they redundantly capture the same signal. The considered method is compared with several tree based methods for feature importance evaluation such as vanilla Gini impurity decrease and more complicated Boruta algorithm. The features are highlighted using a classification algorithm for thick cloud classification based on the marked-up satellite data. Classification accuracy testing based on significant features is performed for different types of surfaces for the set of several single images.

1 INTRODUCTION

Hyperspectral images (HSI) usually acquired from satellites capture light spectra in numerous narrow wavelength ranges. This approach boosts their informational content compared to conventional RGB images. In image processing tasks like cloud detection, monitoring, agriculture and environmental protection (Borzov & Uzilov, 2016), the large volume of HSI data and correlations across neighboring spectral bands require selection of the most informative features.

The existing methods of feature selection assess both feature properties and target variable relationships, covering forward selection and backward elimination methods, exhaustive search, and machine learning-based techniques (Guyon & Elisseeff, 2003). Feature importance evaluation can be done directly by analysing the pair relations between feature and target variable. This approach is generally called filter methods. Correlations between adjacent spectral bands gives folks not only to the idea of using dimentionality reduction algorithms for image compression but also to select relevant features (Myasnikov, 2017) from which PCA is the most popular (Zimichev et al., 2014). This approach along with direct analysis of the correlation matrix doesn't require additional labeling. Such methods as mutual information scores and ANOVA F-value are also very popular filter methods that takes into account labels but doesn't rely on training algorithms. Though these methods are computationally efficient they ignore feature interdependence which can lead to the selection of redundant features that are highly correlated with each other (Zhu et al., 2007). That is why, to select features more effectively, information from the results of HSI classification, in combination with the existing labeling of the source data, is needed. Machine learning (ML) models are widely used to evaluate the importance of features, especially in cases of feature correlations.

Model-specific methods for feature selection integrate the feature selection process directly into the model training phase. This approach offers good balance of performance and computational cost (Saito et al., 2018). There are wrapper and embedded model-specific methods. Wrapper methods identify the optimal subset of features by evaluating their different combinations using a specific

predictive model. Models such as the recursive feature elimination (RFE) select features iteratively, maximizing the performance of the classification model. Although effective in finding optimal subsets, these methods can be computationally demanding (Zubair et al., 2024). Embedded methods integrate feature selection directly into the model training process itself, using mechanisms like regularization penalties to perform selection simultaneously with parameter estimation. Examples of embedded methods include Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge regression (Fira et al., 2025), neural networks with learnable drop layer (JimÉnez-Navarro et al., 2024), sparse principal component analysis (Seghouane et al., 2019), tree-based methods (Tuv et al., 2009).

Decision trees provide an intrinsic feature selection mechanism (Breiman et al., 1984), effectively handling non-linear data and correlated features Kohavi & John (1997), and offering inherent explainability Mishra (2022). Decision tree based models like Random Forests and Gradient Boosting provide feature importance scores based on how much a feature contributes to reducing impurity (e.g., Gini impurity decrease) or variance at each split in the decision trees. More complicated Boruta mathod determine feature relevance by comparing the importance of an original feature with the importance of permutated counterparts known as shadow features and functions as a wrapper algorithm built around the Random Forest classifier (Kursa & Rudnicki, 2010).

In general model specific methods give better results than filter methods but their significant drawback is that the selected set of features is inherently tied to the specific ML algorithm being used (Islam et al., 2022). Stochastic nature of some ML models can lead to reproducibility issues, with variations in feature importance rankings across different model training runs (Vos et al., 2024).

Model-agnostic methods offer greater flexibility and are widely used for interpreting complex "black-box" models. Model-agnostic feature selection methods can be applied to any machine learning model, as they are independent of the model's internal workings (Khan et al., 2025). These methods assess the relevance of features based on their intrinsic properties and their relationship with the target variable, without considering a specific predictive model. Examples of model-agnostic techniques include methods based on eXplainable Artificial Intelligence (XAI) like Permutation Feature Importance (PFI) (Flora et al., 2024) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). These techniques offer flexibility and can be used to compare the importance of features across different models but still affected by feature correlations (Liang et al., 2024) especially PFI (Salih, 2025).

The more sophisticated method of feature importance evaluation is used the more computing time it requires. The more features are taken into account, the more their importance is affected by multicollinearity. This paper is dedicated to the research of the possible approximation for feature importance evaluated from sparse matrix that contains minimal set of values obtained by selection of different feature subsets. We suggest the algorithm of feature importance recovery by means of matrix factorization.

2 Problem statement and method of solution

While high-dimensional HSI data can be rich in information, it also introduces specific obstacles that often reduces the effectiveness of ML models. A family of techniques designed to overcome the curse of dimentionality are commonly addressed through a set of methods known as dimensionality reduction. In this paper the dimentionality reduction is considered as feature selection problem.

The purpose of this work is to propose the algorithm for selecting a limited set of spectral bands and derived features that yields the best prediction quality of dense cloud classification. A dedicated set of labeled images of the HYPERION sensor with a spatial resolution of 30 m and a spectral resolution of 10 m in the spectral range of 400–2500 nm was used. After converting the raw radiance data into reflectance values and eliminating zero channels as well as those corresponding to strong light absorption in water vapour, the selected hyperspectral images (HSI) were organised into a training set. The features of this set comprised reflectance values from spectral channels 8–224, combined with derived characteristics and other indices calculated on a pixel-by-pixel basis. Fig. 1 shows the mean spectral reflectance distribution for cloud and non-cloud pixels, suggesting that a classification algorithm could be developed to differentiate between these pixels based on their

spectral characteristics. In this study, in addition to reflectance values, normalized indices such as NDVI (Huang et al., 2021), NDSI (Jin et al., 2022) and NDWI (Gao, 1996) were used.

We evaluate the importance of features by assess the cloud classification accuracy of a logistic regression model depending on the choice of spectral channels and derived features. The iterative training of ML models with different subset of features is used here to approximate target variable. After training of ML model the corresponding feature importance can be evaluated. By evaluating different sets of features used in the training process, we derive a sparse matrix representing the feature importances for each instance of the trained ML models. This approach yields the following observations:

1. Minimal subset of features is considered because of correlations between them;
2. Training ML models using as many features as possible is computationally expensive;
3. As a result of training ML models, the resulting importance matrix is sparse.

Logistic regression as linear model is strongly affected by features correlation. Decision trees are better but training with large feature subset is still computationally expensive. The resulting importance matrix is sparse, which gives rise to the problem of zero value reconstruction. Such reconstruction is needed to overcome possible feature importance underestimation for correlated features when using Gini impurity decrease as a metric of such evaluation. If we treat feature importance evaluation by approximating from a few values from the sparse matrix one can see that features with high importance ...

So we reformulated it by the following way. The feature importances are modeled as the following matrix decomposition:

$$\hat{T} = PQ^T, \tag{1}$$

where $\hat{T}$ $(n_u \times n_i)$ is the predicted importances corresponding to trained ML model instance u and feature i, P $(n_u \times n_f)$ and Q $(n_i \times n_f)$ are latent factors that capture hidden preferences for features and train instances, respectively. The challenge to the matrix factorization problem is to find P and Q^T. Basically, such an algorithm is going to be used to find latent factors that represent intrinsic feature attributes in a lower dimension (the number n_f of latent factors is choisen beforehand). A learning approach is therefore developed to converge the decomposition results close to the observed importances as much as possible, while ensuring all importance values remain nonnegative. Additionally we introduce the feature bias matrix $\bar{\mu}$ as a correction of (1):

$$\hat{T} = \mu + PQ^T. \tag{2}$$

The feature bias matrix $\bar{\mu}$ supposed to capture the tendency of features to have importance higher (or lower) than the average. So $\bar{\mu}$ measures a trained model tendency to systematically overestimate

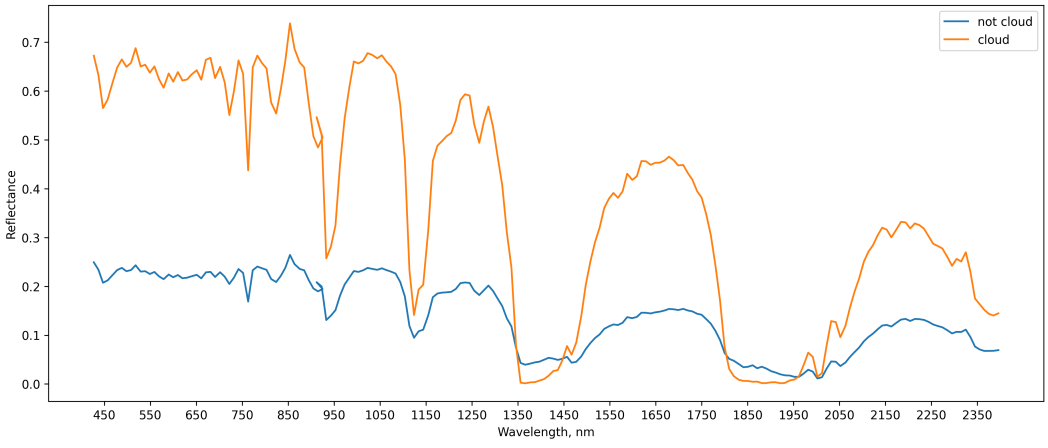

Figure 1: Mean spectral distribution of the reflectance corresponding to cloud and non-cloud pixels

or underestimate feature importances relative to the average across all trained ML model instances. Matrices P, Q and $\bar{\mu}$ can be obtained through a regularized optimization procedure:

$$\sum_u \sum_i \left((T_{u,i} - \hat{T}_{u,i})^2 + \lambda(\mu_{u,i}^2 + ||P_u||^2 + ||Q_i||^2) \right). \tag{3}$$

To bind classification accuracy with feature importance, the correlation between feature importances and classification error is added to the objective function:

$$\sum_u \sum_i \left((T_{u,i} - \hat{T}_{u,i})^2 + R_{u,i} + \lambda(\mu_{u,i}^2 + ||P_u||^2 + ||Q_i||^2) \right), \tag{4}$$

where λ is a regularization parameter, e_u is the classification error of trained ML instance and

$$R_{u,i} = \sum_u \sum_i (e_u - \bar{e}_u)(\hat{T}_{u,i} - \mu_{u,i}) = \sum_u (e_u - \bar{e}_u) \sum_i p_u q_i^T. \tag{5}$$

Here $e_u - \bar{e}_u$ is the deviation of classification error e_u from the average error of classification. The second term of equation (4) accounts for reducing the significance of features in the case of high classification error. If the classification error e_u is below average the importance of features corresponding to the trained ML instance u tends to increase; conversely, if e_u is above average, the corresponding importance values tend to decrease.

Stochastic Gradient Descent used here to solve the problem (2) of matrix factorisation is an optimization algorithm in which the model parameters (in this case, the bias $\bar{\mu}$ and the factor vectors) are repeatedly updated by adding the negative of the gradients calculated with respect to the function (4) being optimized. The algorithm essentially performs the following steps for a given number of iterations:

$$\begin{aligned} \mu_{u,i} &\leftarrow \mu_{u,i} + \gamma(\delta_{ui} - \lambda\mu_{u,i}) \\ p_u &\leftarrow p_u + \gamma\left((\delta_{ui} - 0.5 \cdot e_u) \cdot q_i - \lambda p_u\right) \\ q_i &\leftarrow q_i + \gamma\left((\delta_{ui} - 0.5 \cdot e_u) \cdot p_u - \lambda q_i\right) \end{aligned} \tag{6}$$

where γ is the learning rate and $\delta_{ui} = r_{ui} - \hat{r}_{u,i} = r_{u,i} - (\mu_{u,i} + p_u q_i^T)$ is the error made by the model for the pair (u, i).

3 Results and discussion

The input data for the matrix factorization algorithm is feature importance statistics evaluated by training Decision Tree classifier for different values of hyperparameters. For Decision Tree classifier the decrease of Gini impurity is usually used to assess the feature importance. By considering the accuracy associated with different feature sets, we can calculate the initial correlation between feature importance and training accuracy and sort features according to this value of correlation. Let's refer to this as correlation importance. The most important features include NDWI index and the limited set of features from the NIR and the lower SWIR wavelength range. There is little to no similarity between the feature importance rankings obtained via the Boruta algorithm and those from the present study, with the exception of the consistently high importance assigned to the NDWI feature. This can be explained by the use of shadow features in feature analysis via the Boruta algorithm not used here. However Fig. 2 shows the graph of the random forest model's accuracy versus the selected set of features. By choosing features with higher importance scores (starting from the top of the list in descending order of importance), we obtain models with higher training accuracy.

4 Conclusion

The significant number of channels in HSI, combined with feature multicollinearity, leads to challenges in selecting machine learning models, reducing their accuracy and interpretability. To address this issue for thick cloud classification, decision trees are employed with a selection of a limited number of significant features based on the iterative exclusion algorithm. The proposed method applied

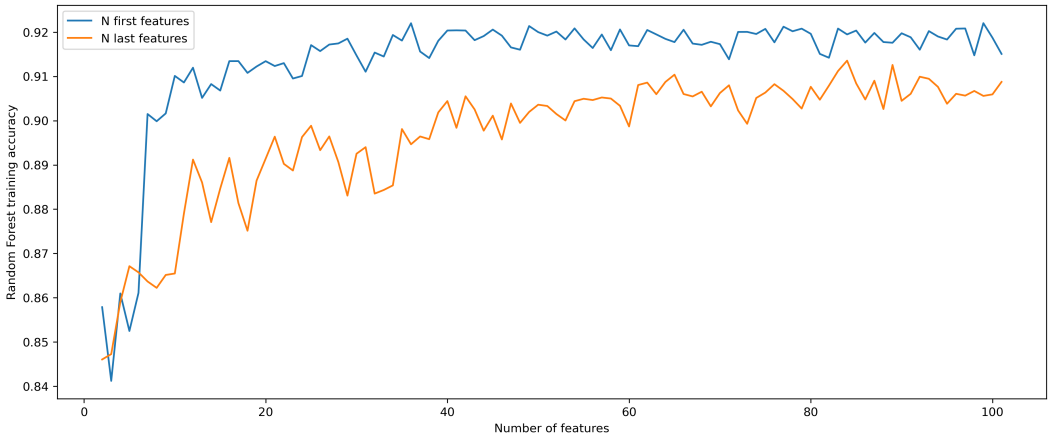

Figure 2: Random Forest training for the first and the last features by their correlation with training accuracy of different Decision Tree classifiers

to find the set of features that yields better prediction accuracy. The most relevant features include NDWI index, limited number of NIR bands and the lower part of SWIR spectrum range bands. Decision Tree model used to assess feature importance effectively handles correlated features avoiding redundancy and can be combined with forward feature selection to achieve more robust statistics. The final set of features selected after applying matrix factorization to overcome sparsity and taking into account the correlation between feature importance and model accuracy. This can be used to construct a classifier for recognizing thick clouds based on their spectral characteristics. Such classifier can be considered a baseline for more complex cloud classification models.

The ambiguity in feature importance rankings produced by traditional methods - such as permutation feature importance, SHAP values, tree-based methods, and linear model coefficients - underscores the need for a more empirically grounded approach. The proposed strategy of iterative feature elimination with feature approximation via matrix decomposition offers an alternative solution to this challenge. This approach not only reveals which features are truly critical for maintaining performance but also accounts for feature interactions and helps identify their optimal subset. One of the way to evaluate feature importance is to use Gini impurity decrease in combination with matrix factorization to overcome sparsity. The point of the present study is to use information about the classification accuracy to evaluate feature importance. This was done by the incorporation of the special correlation term in the objective function for optimization.

Acknowledgments

References

S.M. Borzov and S.B. Uzilov. *J. Comput. Technol.*, 21(1):40 – 48, 2016.

L. Breiman, J. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC., 1984.

Monica Fira, Liviu Goras, and Hariton-Nicolae Costin. Evaluating sparse feature selection methods: A theoretical and empirical perspective. 15(7), 2025. doi: 10.3390/app15073752.

Montgomery L. Flora, Corey K. Potvin, Amy McGovern, and Shawn Handler. A machine learning explainability tutorial for atmospheric sciences. *Artificial Intelligence for the Earth Systems*, 3(1): e230018, 2024. doi: 10.1175/AIES-D-23-0018.1. URL `https://journals.ametsoc.org/view/journals/aies/3/1/AIES-D-23-0018.1.xml`.

B.-C. Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996. doi: https://doi.org/10.1016/S0034-4257(96)00067-3.

Isabelle Guyon and André Elisseeff. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157 – 1182, 2003. doi: 10.1162/153244303322753616.

Sha Huang, Lina Tang, Joseph P. Hupy, Yang Wang, and Guofan Shao. A commentary review on the use of normalized difference vegetation index (ndvi) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, 2021. doi: 10.1007/s11676-020-01155-1.

Md Rashedul Islam, Aklima Akter Lima, Sujoy Chandra Das, M. F. Mridha, Akibur Rahman Prodeep, and Yutaka Watanobe. A comprehensive survey on the process, methods, evaluation, and challenges of feature selection. *IEEE Access*, 10:99595–99632, 2022. doi: 10.1109/ACCESS.2022.3205618.

M J JimÉnez-Navarro, M MartÍnez-Ballesteros, I S Brito, F MartÍnez-Álvarez, and G Asencio-CortÉs. Embedded feature selection for neural networks via learnable drop layer. *Logic Journal of the IGPL*, 33(5):jzae062, 2024. doi: 10.1093/jigpal/jzae062.

Donghyun Jin, Kyeong-Sang Lee, Sungwon Choi, Noh-Hun Seong, Daeseong Jung, Suyoung Sim, Jongho Woo, Uujin Jeon, Yugyeong Byeon, and Kyung-Soo Han. An improvement of snow/cloud discrimination from machine learning using geostationary satellite data. *International Journal of Digital Earth*, 15(1):2355–2375, 2022. doi: 10.1080/17538947.2022.2152886.

Adam Khan, Asad Ali, Jahangir Khan, Fasee Ullah, and Muhammad Faheem. Exploring consistent feature selection for software fault prediction: An xai-based model-agnostic approach. *IEEE Access*, 13:75493–75524, 2025. doi: 10.1109/ACCESS.2025.3558913.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97 (1):273–324, 1997. ISSN 0004-3702. doi: 10.1016/S0004-3702(97)00043-X. URL `https://www.sciencedirect.com/science/article/pii/S000437029700043X`. Relevance.

Miron B. Kursa and Witold R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010. doi: 10.18637/jss.v036.i11.

Annie Liang, Thomas Jemielita, Andy Liaw, Vladimir Svetnik, Lingkang Huang, Richard Baumgartner, and Jason M. Klusowski. Challenges in variable importance ranking under correlation, 2024.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 12 2017. doi: 10.48550/arXiv.1705.07874.

P. Mishra. *Practical explainable AI using python: Artificial Intelligence Model Explanations Using Python-Based Libraries, Extensions, and Frameworks*. New York: Apress, 2022. doi: 10.1007/978-1-4842-7158-2.

E.V. Myasnikov. Hyperspectral image segmentation using dimensionality reduction and classical segmentation approaches. *Comput. Opt.*, 41(4):564–572, 2017. doi: 10.18287/2412-6179-2017-41-4-564-572.

Shota Saito, Shinichi Shirakawa, and Youhei Akimoto. Embedded feature selection using probabilistic model-based optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 1922–1925. Association for Computing Machinery, 2018. doi: 10.1145/3205651.3208227.

Ahmed M. Salih. Re-visiting explainable ai evaluation metrics to identify the most informative features, 2025.

Abd-Krim Seghouane, Navid Shokouhi, and Inge Koch. Sparse principal component analysis with preserved sparsity pattern. *IEEE Transactions on Image Processing*, 28(7):3274–3285, 2019. doi: 10.1109/TIP.2019.2895464.

Eugene Tuv, Alexander Borisov, George Runger, and Kari Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10: 1341–1366, 07 2009. doi: 10.1145/1577069.1755828.

Gideon Vos, Liza van Eijk, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. Stabilizing machine learning for reproducible and explainable results: A novel validation approach to subject-specific insights, 2024.

Zexuan Zhu, Yew-Soon Ong, and Manoranjan Dash. Wrapper–filter feature selection algorithm using a memetic framework. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(1):70–76, 2007. doi: 10.1109/TSMCB.2006.883267.

E.A. Zimichev, N.L. Kazanskiy, and P.G. Serafimovich. Spectral-spatial classification with k-means++ particional clustering. *Comput. Opt.*, 38(2):281–286, 2014. doi: 10.18287/0134-2452-2014-38-2-281-286.

Iqbal Muhammad Zubair, Yung-Seop Lee, and Byunghoon Kim. A new permutation-based method for ranking and selecting group features in multiclass classification. *Applied Sciences*, 14(8), 2024. doi: 10.3390/app14083156.