ENHANCING INTERPRETABLE MACHINE LEARNING FOR EARTH OBSERVATION

Thesis approved by
the Department of Computer Science
University of Kaiserslautern-Landau
for the award of the Doctoral Degree
Doctor of Engineering (Dr.-Ing.)

to

Hiba Najjar

Date of Defense:	16 September 2025
Dean:	Prof. Dr. Christoph Garth
Reviewers:	Prof. Dr. Prof. h.c. Andreas Dengel
	Prof. Dr. Sebastian Vollmer

DE-386

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Hiba Najjar:
ENHANCING INTERPRETABLE MACHINE LEARNING FOR EARTH OBSERVATION

CONTACT INFORMATION:
Email: najjar@rptu.de, hibanajjar998@gmail.com.

To my beloved mother, *Sanaa Zahidi*,
my late father, *Mohammed Najjar*,
and all my brothers and sisters in Gaza.

Abstract

Remote sensing (RS) provides abundant and diverse data for Earth Observation (EO) applications. Machine learning leverages the available data through deep neural networks and specialized architectures. However, increasing model complexity often compromises its interpretability, which is crucial for many EO applications that monitor sensitive human activities or support natural disaster response efforts. This thesis contributes to advancing the interpretability and explainability of complex AI models for various RS applications, with a specific focus on agricultural activities.

Our work employs eXplainable AI (XAI) methods to address two main objective for understanding and improving the model predictions within EO applications. First, we focus on **XAI for Justification** where the model behavior is justified by analyzing how different input features contribute to the outputs. The explanation of individual predictions are leveraged and aggregated to provide a broader understanding of the model's behavior. We apply and evaluate existing model-agnostic and model-specific methods, while also developing new techniques when necessary. We further explore how multi-task learning can further enhance the explainability of predictions.

Second, we apply **XAI for Improvement** based on insights from our prior model justification results. On one hand, we identify the features that are necessary and sufficient for accurate modeling across different contexts. On the other hand, we focus on optimizing the selection and design of vegetation indices, a key component in EO analysis and modeling.

We benchmark our explainability objectives across multiple datasets, covering a range of tasks in EO. We particularly focus on multi-modal datasets, commonly used in EO, to mitigate the research gap regarding the explanation of complex multi-modal networks. The results demonstrate that our approach effectively explains the models by verifying that the model reasoning aligns with expert knowledge. Additionally, our experiments on vegetation indices and the optimization of models through feature reduction yielded promising results, and contributed to enhanced overall model performance and interpretability.

Overall, this thesis provides a thorough examination of the interpretability of ML models under complex modeling scenarios. It leverages various explainability tools and objectives to justify model predictions and improve the modeling strategy and performance. This work contributes an important building block towards more transparent and better performing ML models designed for EO applications.

Acknowledgement

All praise is for Allah, without Whom this thesis would not have come to be, Whose guidance has been constant, and Whose mercy has encompassed me throughout this journey. I am grateful for the clarity of heart He bestowed upon me, allowing me to witness His support and strengthen my faith. And I acknowledge that my praise can never truly encompass the perfection of Allah nor fully express the magnitude of His favors upon me.

I would like to express my gratitude to Prof. Dr. Andreas Dengel for granting me the opportunity to pursue my PhD in his research group, and for his support throughout this journey. It has been a privilege to work with a professor who leads not through pressure and authority, but through kindness, patience, and respect toward all his colleagues and students. I am also grateful for the encouraging and supportive environment you cultivated, which allowed us to thrive and grow seamlessly.

I would also like to extend my gratitude to my mentors, Dr. Diego Arenas and Dr. Marlon Nuske, for their direct supervision throughout my PhD and their invaluable feedback on my work. My thanks also go to all colleagues from the Yield Consortium project. I deeply appreciate the supportive and collaborative environment that each team member helped create.

In last but not least, I would like to express my deepest gratitude to my family for their unwavering support throughout this journey. Not only did they stand by my decision to embark on this path, but they also understood when I considered interrupting it for a noble cause. While the distance between us may have limited physical support, I firmly believe your sincere prayers have been answered, and that this alone has profoundly influenced both the steady progress and completion of this PhD.

Publications

Parts of the research in this thesis, including figures and tables, have already been published in:

JOURNAL

- Günther, A., **Najjar, H.**, & Dengel, A. "Explainable Multi-Modal Learning in Remote Sensing: Challenges and Future Directions". In *IEEE Geoscience and Remote Sensing Letters*. 2024.
- Höhl, A., Obadic, I., Torres, M.Á.F., **Najjar, H.**, Oliveira, D., Akata, Z., Dengel, A. and Zhu, X.X. "Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing". In *IEEE Geoscience and Remote Sensing Magazine*. 2024.
- Mena, F., Pathak, D., **Najjar, H.**, Sanchez, C., Helber, P., Bischke, B., Habelitz, P., Miranda, M., Siddamsetty, J., Nuske, M. and Charfuelan, M. "Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction". In *Remote Sensing of Environment*. 2025.
- **Najjar, H.**, Miranda, M., Nuske, M., Roscher, R., & Dengel, A. "Explainability of Sub-Field Level Crop Yield Prediction using Remote Sensing". In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2025.
- **Najjar, H.**, Pathak, D., Nuske, M., & Dengel, A. "Intrinsic Explainability of Multimodal Learning for Crop Yield Prediction". In *Computers and Electronics in Agriculture*. 2025.

CONFERENCE

- **Najjar, H.**, Helber, P., Bischke, B., Habelitz, P., Sanchez, C., Mena, F., Miranda, M., Pathak, D., Siddamsetty, J., Arenas, D., Vollmer, M., Charfuelan, M., Nuske, M., & Dengel A. "Feature Attribution Methods for Multivariate Time-Series Explainability in Remote Sensing". In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. California, USA, 2023, pp. 5014-5017.
- **Najjar, H.**, Mena, F., Nuske, M., & Dengel, A. "Xai-Guided Enhancement of Vegetation Indices for Crop Mapping". In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. Athens, Greece, 2024, pp. 4140-4144.
- **Najjar, H.**, Alshbib, B., & Dengel, A. "Can Multitask Learning Enhance Model Intrinsic Interpretability?". In *47th DAGM German Conference, DAGM GCPR 2025, Freiburg, Germany. Proceedings.*.

WORKSHOP

- **Najjar, H.**, Mena, F., Nuske, M., Dengel, A. "Xai-Guided Enhancement of Vegetation Indices for Crop Mapping". In *Twenty-Fourth AAAI Conference on Artificial Intelligence 2024 Explainable machine learning for sciences workshop*. Vancouver, Canada, 2024.
- **Najjar, H.**, Nuske, M., Dengel, A. "Data-Centric Machine Learning for Earth Observation: Necessary and Sufficient Features". In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD. Cham: Springer Nature Switzerland*. Vilnius, Lithuania, 2024.

Contents

List of Figures

List of Tables

List of Acronyms

R^2	coefficient of determination. 46, 49, 50, 53, 62–65, 75–78, 91, 106, 108, 127, 129, 133, 159
1D-CNN	1-Dimensional convolutional neural network. 73, 74, 76, 78–80
AI	Artificial Intelligence. 5, 6, 10, 14, 15, 17, 20, 21, 23, 31, 149
ALE	Accumulated Local Effects. 25
ALSTM	attention-based long short-term memory. 69, 73, 78–80
AR	Attention Rollout. 83, 84, 87–92, 95–97, 150, 167, 168
CAM	Class Activation Mapping. 27–29, 93
CNN	convolutional neural network. 18, 24, 27, 29
DEM	digital elevation map. 43–45, 60, 71, 73, 80, 94, 102, 108, 114, 117, 118
DL	Deep Learning. 3–5, 7, 18, 28, 31, 48, 137–139, 149
DT	decision tree. 17
EO	Earth Observation. 3, 7, 8, 10, 13, 17–19, 23, 29, 31, 35, 76, 81, 123, 124, 126, 137, 138, 149–151, 153, 157
EVI	Enhanced Vegetation Index. 137
FAO	Food and Agriculture Organization. 15
GA	Generic Attention. 83–85, 87–90, 95, 97, 150, 167, 168
GB	Guided Backprop. 128, 130–136
Grad-CAM	Gradient-weighted CAM. 24, 28, 29, 93
GRU	gated recurrent Unit. 127, 139, 140
I*G	Input $\times$ Gradients. 27, 37
ICE	Individual Conditional Expectation. 25
IG	Integrated Gradients. 27, 38, 39
IoU	intersection over union. 106, 108, 160
k-NN	k-nearest neighbor. 17
L-TAE	lightweight-TAE. 127
LiDAR	Light Detection And Ranging. 13, 14, 20
LIME	Local Interpretable Model-agnostic Explanation. 26, 29, 38, 39

LRP Layer-wise Relevance Propagation. 24, 27
LSTM long short-term memory neural network. 45, 69, 72, 78–80, 94, 127, 128
LULC Land Use and Land Cover. 103, 108, 113, 114, 116–118

MAE mean absolute error. 45, 46, 75–78, 106, 110, 118–120, 159
MHA multi-head self-attention. 74
ML Machine Learning. 3–5, 7, 8, 10, 17, 19, 21, 23, 28, 29, 31, 42, 48, 76, 123, 124, 126, 128, 138, 149, 157, 161
MLP multilayer perceptron. 4, 5, 71, 72, 80, 93, 105, 127
MSE mean squared error. 105

n-NDVI narrow Normalized Difference Vegetation Index. 143, 144
n-NIR narrow Near-InfraRed. 139, 140, 143
NDMI Normalized Difference Water Index. 143, 144
NDRE Normalized Difference Red Edge. 142–145
NDVI Normalized Difference Vegetation Index. 54, 137, 143, 144
NIR Near-InfraRed. 15, 61, 139, 140, 142, 143
NN neural network. 17

OA overall accuracy. 139, 140, 143–145

p.p percentage points. 49
PCA Principal Component Analysis. 64, 66
PDP Partial Dependence Plot. 25
PM Particulate Matter. 127

RADAR RAdio Detection And Ranging. 13
RE Red-Edge. 61, 130, 137, 139, 140, 142, 144, 145
RF random forest. 4, 17, 28–30
RGB Red-Green-Blue. 6, 20, 54
RMSE root mean square error. 46, 75–78, 80, 81, 159
RNN recurrent neural network. 18, 72–74, 127, 128
ROAR RemOve And Retrain. 125, 133
RS Remote Sensing. 3, 4, 6–11, 14, 17, 18, 20, 24, 28–31, 35, 68, 74, 101, 102, 105, 112, 119, 125, 149, 150

S1 Sentinel-1. 103, 126
S2 Sentinel-2. 43, 44, 48, 53, 54, 56, 60, 65, 80, 103, 126, 127, 133, 137–139, 142, 145
SAR synthetic aperture radar. 14, 15, 19, 20, 29, 101, 103, 108, 150
SAVI Soil-Adjusted Vegetation Index. 137
SCL scene classification layer. 43
SG-SQ SmoothGrad-Squared. 128, 133–135
SGD Sustainable Development Goal. 6

SHAP SHapley Additive exPlanations. 26, 28, 30, 68
SRTM Shuttle Radar Topography Mission. 43
SVM support vector machine. 4, 17, 29, 30, 124, 125
SVS Shapley Value Sampling. 38, 39, 51–53, 66, 67, 87–89, 92–97, 128–136, 139, 149–153, 163, 167, 168, 170, 172
SWIR Short-Wave InfraRed. 13, 55, 130, 133, 137, 139, 140, 142–145

TAE temporal attention encoder. 127
TempCNN temporal convolutional neural networks. 127, 130
twi Topographic Wetness Index. 43

USDA United States Department of Agriculture. 15

VAR VarGrad. 128, 133–135
VI vegetation index. 48, 49, 54, 55, 137, 138, 140, 142–145, 151, 153

WMA Weighted Modality Activation. 93, 94, 97, 150, 152, 153, 170, 172

XAI eXplainable AI. 6, 7, 10, 20–24, 28–31, 35, 36, 55, 100, 124, 128, 138, 145, 149–153, 157, 163

PartI

INTRODUCTION

1 Introduction

1.1 MOTIVATION

1.1.1 Important role of Machine Learning in Earth Observation

Abundance of satellites and satellite data

As of May 1st, 2025, more than 8,000 active satellites are orbiting Earth across low, medium and geosynchronous orbits, as reported by the satellites tracking website *Orbit Now* [1]. Approximately 90% of these satellites are designed for communications and Earth Observation (EO) missions [2]. EO satellites continuously collect and store large amounts of data, much of which is transmitted to ground servers for downstream tasks such as weather forecasting, agricultural monitoring, urban planning, and environmental change tracking. While commercial and military satellite data are typically restricted from public use or available only through commercial licensing, government-operated and open-data satellites are made fully open-access, serving as invaluable resources for interdisciplinary research. However, satellite imagery is far more complex than ordinary photographs taken with our personal cameras, due to factors such as spectral resolution, radiometric calibration, and atmospheric interference.

Machine-based processing of satellite data

The diversity and large volume of Remote Sensing (RS) data necessitates machine-based processing. Machine Learning (ML) and Deep Learning (DL) techniques have been increasingly used in recent years to optimize the processing of satellite-derived data and enhance model performance for various downstream tasks [312]. This synergy is a perfect match: while ML thrives on large datasets, RS is inherently data-abundant. The continuous growing research in computer vision and natural language processing is inspiring practitioners in RS and EO, encouraging the development of methods to maximize the potential of the available satellite-derived data.

1.1.2 Interpretability between Classical ML and Advanced DL

Impressive performance of deep learning

DL has a history of impressive achievements in the last two decades, spanning computer vision, reinforcement learning, generative models, and the ongoing rise of large language models. And yet, we might just be scratching the surface of the wonders DL can achieve. The potential of DL has also attracted researchers in RS, eager to claim their share of the cake. As the unique characteristics of satellite data raise new challenges for modeling tasks, deep networks is being leveraged to address these scientific challenges.

Lack of interpretability in deep networks

Despite their glamorous history of breaking records in model perfor-

[1] https://orbit.ing-now.com/
[2] https://www.ucs.org/resources/satellite-database

mance, sometimes even beyond human-level accuracy, deep networks are often complex architectures that behave like black boxes. Their depth, width, and perplexing structures operate at a high level of abstraction. It is no longer easily possible for humans to track the reasoning process and patterns learned by the model. The complexity of deep networks has enabled significant gains in prediction accuracy, yet it often comes at the cost of model transparency and interpretability.

Examples of interpretable ML models

To illustrate the interpretability gap between classical ML models and deep networks in DL, let us consider some models generally recognized as easily understandable [141, 88, 116]. *Linear regression models* learn weights that directly indicate how much each feature contributes, positively or negatively, to the prediction. A *decision tree* has a graphical structure and typically uses only a subset of the input features (rather than all). These aspects, among other, provides both a global understanding of the tree, by examining the splitting rules it has learned, and a local explanation of individual predictions, by tracing the path and rules a sample follows to reach its prediction leaf. A linear *support vector machine (SVM)* can be interpreted through its decision boundary, represented by a hyperplane in the feature space. The hyperplane's weights reflect the significance of each feature in class separation, while the support vectors (i.e. data points closest to the decision boundary) offer a clear visual justification for the placement of this hyperplane.

Decrease of interpretability in more complex models

In contrast, increasing the network's depth and/or complexity progressively challenges interpretability. *Random forests (RFs)* combine multiple decision trees and aggregate their predictions, averaging them for regression tasks or taking the majority vote for classification tasks. Interpreting each tree and understanding how they collectively contribute to predictions becomes increasingly difficult as the number of sub-trees grows. *Kernel-based SVMs*, which map input samples into a new space (i.e. the kernel space) to define the decision boundary, also reduce interpretability by introducing non-linearities between the input and boundary spaces. *Multilayer perceptrons (MLPs)*, which behave like linear regressors when containing only a single layer, quickly lose interpretability as the number of layers (depth) and neurons (width) increases, compounded by the non-linear activation functions applied between layers. Similarly, *neural networks* extend MLPs with complex architectures composed of intricately entangled blocks, further intensifying the interpretability challenge.

Why care about interpretability alongside accuracy?

If model performance improves, why should we still be concerned about interpretability? There are indeed several reasons to strive for a fair balance between performance and interpretability. In practice, DL models are trained to be deployed in larger systems for a wide range of applications. During the deployment phase, new data samples are fed to the model, and its capacity to generalize to this new data is crucial to ensure that the model maintains similar performance to what was achieved on the training samples. A key factor contributing to the model's generalizability is the correctness of the reasoning process it has learned to map inputs to their corresponding outputs. Interpreting the model allows us to verify whether this reasoning process aligns with domain knowledge or common sense. When combined with performance metrics, interpretability ensures that the model is making the right decisions for

the right reasons. If inconsistencies with domain knowledge are found, this suggests that the model is susceptible to failure when processing new, unseen data. It also provides valuable feedback for ML practitioners to refine the model, ensuring that its reasoning is more robust and in line with the correct understanding required for the task.

The end-user right for a justification

Interpretability is also essential for the end user of the Artificial Intelligence (AI) system: users should be able to request justifications for model predictions at any time. Black-box models, by nature, fail to provide such justifications. On the other hand, interpretability ensures that the model can meet these requests and comply with legal requirements. In fact, the European AI Act, effective as of August 2024, took years of preparation to establish a regulatory framework to protect the constitutional and fundamental rights of European citizens with respect to AI systems. The first clause in **Article 86: Right to Explanation of Individual Decision-Making** states:

> *Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system [...] shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.*

1.1.3 Role of Interpretability in Earth Observation

Critical applications in Earth Observation

While high-risk systems are particularly subject to strict regulations to ensure a safe, ethical and trustworthy deployment and usage of AI, RS data is also essentially used in a range of critical applications. Governments use satellite data for border surveillance and to monitor activities in conflict zones, such as illegal fishing or mining. Environmental monitoring also benefits greatly from RS data, enabling the detection of oil spills in oceans, tracking extreme weather events like heatwaves and droughts, and managing natural disasters, including floods, earthquakes, and wildfires. By identifying and monitoring these events, satellite-derived data helps coordinate relief efforts and supports decision-making in response to crises.

Agriculture is another domain where RS data plays a crucial role. Satellite data helps track threats to crop yields from factors such as droughts, floods, and outbreaks of disease and pests. It also assists in monitoring crop types and quantifying agricultural production, enabling better planning for resource allocation, export and import strategies, and market stabilization. Most importantly, RS data supports efforts to ensure food security —an issue of growing global concern, as highlighted by its second rank in the Sustainable Development Goals (SGDs):

> ***Sustainable Development Goal 2***
> *End hunger, achieve food security and improved nutrition and promote sustainable agriculture.*

1.1.4 Limitations of Existing Explainability Techniques

Explainability between computer vision and earth observation tasks

EXplainable AI (XAI) research has seen exponential growth in the past two decades. The methods developed have primarily focused on computer vision and natural language tasks, often involving Red-Green-Blue (RGB) images and sequential textual data, which are typically processed by unimodal models (i.e., models that handle a single input modality). However, transferring these methods to RS presents multiple challenges due to the unique characteristics of satellite-derived data. Satellite imagery differs significantly from natural images in terms of spectral, spatial, and temporal resolutions, making it distinct from typical computer vision benchmark datasets. Moreover, the satellite-derived data is inherently mutlimodal and frequently necessitate advanced data fusion techniques and sophisticated model architectures. These distinctive aspects make it challenging to directly transfer existing interpretability and explainability techniques from traditional domains to EO applications.

This thesis aims to bridge the gap between the existing literature on interpretable DL and RS applications. We evaluate and adapt existing methods to meet the complex requirements and modeling challenges posed by satellite-derived data. In an effort to close the ML loop, we also explore how the outcomes of these explanations can be leveraged to improve feature engineering and enhance modeling strategies.

1.2 RESEARCH QUESTION

The main objective of this thesis is to enhance the interpretability of deep networks deployed for EO applications. Building upon the existing body of research in XAI and the wide range of interpretability tools it offers, this work addresses the following main research question:

How can the interpretability of deep networks be enhanced for EO applications?

To answer this question, we consider two main objectives of model interpretability:

1. **Justification:** Model interpretations provide both local and global explanations. Local explanations address individual predictions, offering insight into why a specific decision was made. Global explanations, on the other hand, uncovers the overall functioning of the model. Both levels facilitate the justification of the model behavior, enabling the validation of its reasoning process, and upholding the right to explanation for end users.

2. **Improvement:** In an attempt to close the loop between the data processing, modeling, and interpreting stages, explanation outcomes can be harnessed to support the two initial steps, enhancing the feature engineering and modeling strategies.

In this context, and within the framework of the two interpretability goals described above, we address the main research question through the following three sub-questions and their corresponding objectives:

1. **Question:** How can existing interpretability techniques explain ML models designed for EO?

   Goals: Evaluate existing techniques on EO models. Identify and apply a reliable method to justify the model predictions. Validate the model's reasoning process against domain knowledge.

2. **Question:** Can the multimodal nature of RS data be leveraged for model interpretability?

   Goals: Adjust existing techniques for multimodal learning scenarios. Design new methods for intrinsic interpretation. Leverage multi-task learning to justify model predictions.

3. **Question:** How can explanation results contribute to improving the data processing and modeling phase?

   Goal: Identify necessary and sufficient features and modalities to optimize the size of the input space and preserve model performance. Use explanation results to efficiently select satellite-derived indices for the EO task under study.

1.3 CONTRIBUTIONS

This thesis is mainly contributing to the interpretability of ML models designed for EO applications. Our research acknowledges the unique characteristics of RS data and their influence on modeling strategies. Accordingly, we progressively integrate interpretability tools into this context, adapting them to align with the specific requirements of RS data while enhancing their utility for EO tasks. Below, we outline the key contributions of this thesis:

- We systematically review the existing body of literature explaining ML models across different EO applications. We analyse leading trends in the usage of explainability techniques among EO practitioners, and identify key methodological limitations. This work has been published in:

  ➔ Höhl, A., Obadic, I., Torres, M.Á.F., **Najjar, H.**, Oliveira, D., Akata, Z., Dengel, A. and Zhu, X.X. "Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing". In *IEEE Geoscience and Remote Sensing Magazine*. 2024.
  ➔ Günther, A., **Najjar, H.**, & Dengel, A. "Explainable Multi-Modal Learning in Remote Sensing: Challenges and Future Directions". In *IEEE Geoscience and Remote Sensing Letters*. 2024.

- We evaluate existing interpretability tools on a challenging RS task, focusing on feature attribution methods. A diverse set of techniques is selected and applied to the crop yield prediction task, including multiple input modalities. Both qualitative and quantitative evaluations are conducted, with the best-performing method used to explain the

model's reasoning and validate it against expert knowledge from the agronomic domain. This work has been published in:

➔ **Najjar, H.**, Helber, P., Bischke, B., Habelitz, P., Sanchez, C., Mena, F., Miranda, M., Pathak, D., Siddamsetty, J., Arenas, D., Vollmer, M., Charfuelan, M., Nuske, M., & Dengel A. "Feature Attribution Methods for Multivariate Time-Series Explainability in Remote Sensing". In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. California, USA, 2023, pp. 5014-5017.

➔ **Najjar, H.**, Miranda, M., Nuske, M., Roscher, R., & Dengel, A. "Explainability of Sub-Field Level Crop Yield Prediction using Remote Sensing". In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2025.

- We increase the model complexity by transitioning from early modality fusion to an intermediate fusion approach, allowing more efficient encoding of each modality. Given the limited focus on interpretability within multimodal learning frameworks in existing literature, we propose adaptations of existing explainability techniques for this task and conduct extensive experiments to evaluate their performance. This work has been published in:

  ➔ Mena, F., Pathak, D., **Najjar, H.**, Sanchez, C., Helber, P., Bischke, B., Habelitz, P., Miranda, M., Siddamsetty, J., Nuske, M. and Charfuelan, M. "Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction". In *Remote Sensing of Environment*. 2025.

  ➔ **Najjar, H.**, Pathak, D., Nuske, M., & Dengel, A. "Intrinsic Explainability of Multimodal Learning for Crop Yield Prediction". In *Computers and Electronics in Agriculture*. [under review]

- We leverage the availability of various modalities in typical RS applications to explore the multi-task learning framework. We use additional input modalities as auxiliary task, and leverage this setup to explain the main task. Specifically, we show how the predictions on the main task can be explained through the auxiliary tasks, revealing the model behavior as influenced by the complementary variables it predicts. This work has been published in:

  ➔ **Najjar, H.**, Alshbib, B., & Dengel, A. "Can Multitask Learning Enhance Model Intrinsic Interpretability?". In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [under review]

- We bridge the gap between data preparation and model interpretation by exploiting explanation results to refine the feature engineering process. Because of the several satellite-derived modalities usually used and the large number of features they include, we use the feature attribution results to iteratively identify a sufficient and necessary set of features. The framework proposed achieves an optimized definition of the input space while maintaining a good model performance. This work has been published in:

➔ **Najjar, H.**, Nuske, M., Dengel, A. "Data-Centric Machine Learning for Earth Observation: Necessary and Sufficient Features". In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD. Cham: Springer Nature Switzerland*. Vilnius, Lithuania, 2024.

- We address RS applications that often rely exclusively on indices for modeling. Leveraging explanation results, we guide the selection of appropriate variables from the extensive library of indices in the literature. This approach improves both model performance and the interpretability of the input space. This work has been published in:

  ➔ **Najjar, H.**, Mena, F., Nuske, M., & Dengel, A. "Xai-Guided Enhancement of Vegetation Indices for Crop Mapping". In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. Athens, Greece, 2024, pp. 4140-4144.

1.4 THESIS STRUCTURE

Following the previously introduced motivation (Section 1.1), research questions (Section 1.2), and contributions of the thesis (Section 1.3), we will elaborate on the background of this work in the next chapter. Specifically, it describes the role of AI and ML in EO, examines the value of XAI for EO applications, introduces explainability objectives and feature attribution techniques, and presents current trends in the usage of explainability techniques among EO practitioners along with key methodological limitations.

The remaining chapters of the thesis are divided into two main parts, which align with the two main objectives of model interpretability which we will address: justification and improvement.

Part II - XAI for Justification

Chapter 3 evaluate existing interpretability tools on the task of crop yield prediction, focusing on feature attribution methods and multimodal learning networks. Post-hoc methods are investigated in Section 3.2, under the early fusion setup of the input modalities. Ante-hoc (intrinsic) techniques are explored in Section 3.3, within an intermediate fusion framework.

Chapter 4 investigates the application of multi-task learning in RS by exploiting the inherent availability of multiple data modalities. The additional input modalities are used as auxiliary tasks to enhance the primary task performance and provide intrinsic model interpretability.

Part III - XAI for Improvement

Chapter 5 bridges the gap between data preparation and model interpretation by employing explanation results to refine the feature engineering process. Feature attribution results are used to iteratively identify a sufficient and necessary set of features to optimize both model accuracy and input space efficiency.

Chapter 6 introduces an explanation-driven framework for optimizing

vegetation index selection in RS applications. The framework uses feature attribution results to identify the most informative indices from existing literature.

The thesis is concluded in Part IV with an overall summary and future work.

Background

2

2.1 ARTIFICIAL INTELLIGENCE FOR EARTH OBSERVATION

2.1.1 Earth Observation Data

Satellite missions

Satellites are launched for a variety of missions, including communications, earth surface observation, navigation, space science, and technology development [167, 297, 93]. Specifically, EO satellites are designed to monitor and collect data about the Earth's surface, atmosphere, and oceans, serving applications in environmental management, agriculture, meteorology, and security. Notable EO missions include Landsat, MODIS, Copernicus Sentinel, and TerraSAR-X [34].

Passive sensors

Active sensors

We can distinguish between two main types of sensors used in EO satellites: passive and active. Passive sensors measure the energy reflected or emitted by the Earth's surface: optical and infrared sensors detect reflected sunlight, while thermal infrared sensors measure the Earth's emitted thermal radiation. Active satellites, such as RAdio Detection And Ranging (RADAR) and Light Detection And Ranging (LiDAR), emit their own signals and then record the reflected signals to capture data about the Earth's surface. These sensors are not dependent on natural light and can capture information even in dark or cloudy conditions, making them particularly useful in such challenging environments. In the following, we describe the main characteristics of the data captured by each type of sensors:

Passive satellites

Spectral categories

Active satellites

Spectral resolution. The wavelength range of passive satellites typically extends from the visible and near-infrared regions to the Short-Wave InfraRed (SWIR). Depending on the number of spectral bands used during the imaging process, we distinguish between three main categories of passive satellites: *panchromatic*, with a single broad spectral band, *multispectral*, with few spectral bands within narrow wavelength ranges, and *hyperspectral*, which captures 100 or more contiguous bands for precise spectrum information. In contrast, active satellites have typically a limited spectral resolution, which consists of characteristics of the emitted and reflected signals rather than a range of wavelengths.

Passive satellites

Active satellites

Spatial resolution. Passive satellites have a wide range of spatial resolutions specific to their missions. For instance, Landsat has moderate spatial resolution at 30 meters, Sentinel-2 achieves better resolution, capturing some bands at 10 meters, while higher-resolution systems like WorldView-3 can offer very fine spatial resolution, up to 31 cm [93]. Active satellites also vary in spatial resolution based on sensor power and mission objectives: synthetic aperture radar (SAR) satellites like Sentinel-1 generally have spatial resolutions between 10–20 meters, while LiDAR

systems onboard satellites often deliver resolutions in the meter range, with some exceptional satellites, such as ICESat-2, which can achieve highly detailed vertical precision of up to 2 centimeters for elevation data [298].

Passive satellites

Temporal resolution. Defined as the frequency of revisiting the same geolocation, the temporal resolution of passive satellites is influenced by their orbit, which must be sun-synchronous to ensure consistent lighting conditions, and the width of their sensor's field of view. For instance, Landsat has a 16-days revisit cycle because of its narrow width of 185km. MODIS, in contrast, ensures a daily global coverage thanks to its very wide 2,330 km width but at low spatial resolution. Sentinel-2 is composed of two satellites working in tandem to optimize its revisiting frequency, which reaches 5-days at the equator, with a 290 km swath width [252].

Active satellites

Active satellites, independent of sunlight, can achieve higher temporal resolutions. Certain missions can even offer revisit intervals as short as 1–3 days in areas of interest, defined as high priority zone.

2.1.2 *Earth Observation for Agriculture*

Satellite data find applications across diverse fields, including weather forecasting, environmental monitoring, agriculture, urban planning, and defense surveillance. Particularly in agriculture, RS data hold great potential to transform agricultural practices and support decision-making processes for various stakeholders, including farmers, agribusinesses, and organizations involved in the agricultural sector [148, 326]. Alongside AI and other complementary technologies in robotics, RS data plays a central role in smart farming, which aims at enhancing the efficacy and profitability of crop production while minimizing inputs in energy and agrochemicals [143]. Among the many agricultural activities supported by satellite data, *crop identification* and *crop production* are particularly prominent and are discussed in detail below.

RS for Agriculture

Smart farming

Nomenclature

Crop identification, also known as **crop classification** or **crop mapping**, enables the generation of agricultural maps on regional and larger scales, facilitating efficient crop management. Reliable and up-to-date crop maps play an essential role in guiding resource allocation efforts. They support balancing water availability with irrigation schemes, adapting crops to climate change, ensuring crop varieties meet market demand, monitoring land use changes over time, and promoting sustainable practices through crop rotations and intercropping, among other strategies [9]. The task of crop classification can be conducted using both active and passive satellites. For instance, SAR data help identify characteristics of the canopy structure, soil surface and water content, while optical satellites are more reliable for capturing spectral properties such as pigments, flower colouring and Near-InfraRed (NIR) reflectance strength [237]. These complementary properties enable AI systems to effectively distinguish and classify crop types.

Goals and applications

Satellite types used for crop classification

Nomenclature

Crop production, also known as **yield estimation** or **yield prediction**, is the task of estimating the amount of crop in tons per hectare (t/ha). Depending on the inputs, this task predicts either the potential future

Types of target values

yield, defined as the maximum yield achievable under optimal and controlled conditions for a specific crop cultivar, or the actual harvested yield at the end of the growth season. On one hand, forecasts of crop yield can support local efforts to enhance agricultural profitability and inform regional strategies. This can be achieved by comparing estimated yields with current and future demands, subsequently adjusting import and export plans, informing market strategies, and contributing to international efforts for ensuring global food security. On the other hand, post-harvest yield estimates play a key role in completing historical yield records and addressing potential data gaps. For example, such yield estimates are particularly valuable for government agencies such as the United States Department of Agriculture (USDA) and Eurostat, as well as organizations like the Food and Agriculture Organization (FAO), to publish agricultural yield data at regional and national levels. Additionally, universities with agricultural research programs rely on historical yield records to assess the impact of weather and other factors on crop yields. Satellite data providers also benefit from such yield estimates, as they can incorporate them into agricultural data offerings alongside primary satellite imagery.

Yield forecast

Post-harvest yield prediction

Table 2.1 summarizes the features extracted from various satellite missions and highlights their applications in the crop identification and yield prediction literature.

2.1.3 From Classical to Advanced Machine Learning

As early as in the 1960s, the first computer-based processing systems for RS applications emerged [175, 92]. Building on these early implementations, researchers explored in detail the potential of this field, discussing prospective uses and applications of AI in RS [174, 78]. In practice, the earliest machine-based systems applied in this domain were rule-based systems, manually configured by experts based on domain knowledge, commonly referred to as *expert systems* [114, 339, 159, 303]. Subsequently, various traditional ML models were adopted. Among these were k-nearest neighbors (k-NNs) [305, 250, 83], decision trees (DTs) [216, 121, 90, 56], maximum likelihood [216, 121, 90],mostly used for classification tasks, and linear regressors [89, 247, 193, 54, 28, 152] and regression trees [214, 223, 57, 180], primarily for regression tasks. To address the challenges posed by high-dimensional satellite data, such as hyperspectral imagery, and to better model complex, nonlinear relationships, neural networks (NNs) became widely used [128, 327, 70, 25, 18, 265], followed by the SVMs from the late 1990s onwards [242, 20, 61, 37, 38].

Computer-based processing

Expert systems

Traditional ML

Neural networks

Support vector machines

On account of the increasing availability of satellite data and advancements in algorithms, novel techniques were adopted in the 2000s. Notably, ensemble-based approaches gained significant attention in EO applications, due to their robustness in handling noise and mitigating overfitting issues [40, 41, 104]. Prominent methods included bagging [36, 40, 107], boosting [36, 40, 107, 23, 41, 325], and RFs [239, 107, 177, 41, 325]. Other specialized models were also designed to address specific tasks, including target detection [264, 292, 185], change monitoring [197, 322, 153], and anomaly detection [172, 24, 71].

Ensemble-based modeling

Specialized models

Tab. 2.1.: Satellite missions used in crop mapping and yield estimation. (The list in not exhaustive.)

Sensor	Missions	Agricultural Task	Satellite-derived features	Study
Optical	Landsat-5-7-8	Corn/soybean crop mapping	NIR, SWIR1, SWIR2 bands and GCVI index	Claverie et al. [51]
	Landsat-8	Cotton yield prediction	NDVI, SR, NIR, Green-NDVI, GI, WI, and SBI	Haghverdi et al. [119]
	MODIS	Soybeans/strawberries yield prediction	surface reflectance and land surface temperature	Gastli et al. [99]
	Sentinel-2	Rice yield prediction	spectral bands and cloud mask	Son et al. [279]
	Sentinel-2	Corn yield prediction	visible (B2,B3,B4), near infrared (B8), red-edge (B5,B6,B7) and shortwave infrared (B11,B12) bands, and Green-NDVI, NDRE, and NDWI indices	Desloires et al. [59]
	Sentinel-2	Crop mapping	visible (B2,B3,B4), near infrared (B8) bands and NDVI index	Belgiu and Csillik [27]
	Pléiades	Winter cereals/ oilseed rape crop mapping	visible and near infrared bands	Vaudour et al. [317]
	WorldView-2	Wheat yield prediction	8-band multispectral and a panchromatic images	Tattaris et al. [296]
SAR/ Li-DAR	TerraSAR-X, RADARSAT-2	Corn/soybean crop mapping	linear polarizations (HH, VV, VH/HV)	McNairn et al. [208]
	ALOS/ POLSAR	Crop mapping of beans, beets, grasses, maize, potato and wheat	HH polarizations	Sonobe et al. [280]
	Sentinel-1A	Corn/soybean/wheat yield prediction	VH, VV, VH/VV, and radar indices	Hashemi et al. [122]
	Radarsat-2	Rice yield prediction	HH, HV polarizations	Zhang et al. [346]

Deep learning

The 2010s marked a paradigm shift with the rise of DL, which transformed conventional image processing approaches, particularly through the adoption of convolutional neural networks (CNNs). Although first introduced in 1989 [178], CNNs gained widespread recognition in 2012 with the AlexNet architecture [169]. These networks significantly optimized the feature engineering techniques by efficiently extracting features from raw data, overcoming the requirement of domain expertise for manual feature design [184]. In RS, CNNs were applied to various tasks such as land cover classification [266, 171, 137, 342], object/cloud detection [49, 58, 273, 270, 335], and image segmentation [157, 48, 291, 235].

Convolutional neural networks

Recurrent neural networks

Temporal data was also blessed by the introduction of recurrent neural networks (RNNs), which proved particularly effective for forecasting tasks in weather prediction [263, 274, 6], agriculture monitoring [334, 207, 103], and object/cloud detection [49, 58, 273, 335, 270] tasks. The adoption of transfer learning further enhanced RS tasks by leveraging models pre-trained on different geographical regions or general-purpose computer vision tasks. This approach improved model performance and mitigated limitations in labeled satellite data [323, 262, 245, 206, 338, 347].

Transfer learning

Unsupervised learning, Transformers, Generative AI

Since the late 2010s until now, advancements in DL have been increasingly applied to EO applications, with a focus on addressing challenges such as reducing reliance on labeled data through semi- and self-supervised learning [135, 133, 156, 182, 319, 324], analyzing long time series with Transformer models [44, 43, 80, 343, 205, 64], and using generative models for certain tasks, including image translation [313, 183, 166], multisensor data fusion [45, 331, 66] and spatial resolution enhancement [149, 309, 310].

2.1.4 *Multimodal Learning*

Satellite data diversity

The diverse types of sensors launched into space over recent decades collect data that vary significantly in their characteristics. As a result, in practice, data from multiple sources are used together to achieve a specific task. Within the context of ML, the field of multimodal learning focuses on optimizing architectural design and computational approaches to facilitate the fusion of multiple modalities, thereby enhancing model performance. In this thesis, the term *multimodal learning* refers to any learning process that integrates data from multiple sources, combining them at some stage to produce final predictions. For EO tasks, these modalities may consist exclusively of satellite-derived data or may also include metadata or in-situ sensing data relevant to the task. Nevertheless, surface reflectance remains the most commonly used type of satellite data in multimodal learning, often complemented by SAR data [212].

Multimodal learning

Fusion types

We distinguish between four primary types of multimodal learning techniques based on the stage at which the fusion of different modalities occurs: early, intermediate, late, and hybrid fusion. In **early fusion**, the modalities are combined before being supplied to the ML model. For spatial and/or temporal data, this approach may require additional pre-processing steps to align the spatial and temporal resolutions of the modalities. However, models originally designed for a single modal-

Early fusion

ity can often be applied directly in this context, and typically require minimal to no modifications. In **intermediate fusion**, the modalities are initially processed individually by separate networks, commonly referred to as encoders, and are often projected into a shared representation space. These representations are then combined and processed by an additional network, referred to as the task head. The intermediate fusion can be implemented using various techniques, including simple concatenation, weighted average, element-wise product, maximum pooling, attention-based mechanisms, or gated mechanism [321, 15, 212]. In **late fusion**, the modalities are also processed independently, but their respective encoders are trained directly on the target outputs. The individual predictions are subsequently combined to produce the final output using techniques such as maximum pooling, majority voting, or weighted averaging [321]. Additionally, uncertainty estimation can be employed to prioritize the most confident predictions [120, 332]. Finally, the **hybrid fusion** integrates elements of the aforementioned fusion strategies. For instance, modalities can be processed using a two-branch architecture, where one branch employs early fusion while the other follows an intermediate fusion approach. The outputs of these branches are then combined using techniques from the late fusion paradigm to produce the final prediction.

Intermediate fusion

Late fusion

Hybrid fusion

Comparison of fusion techniques

Intermediate fusion outperformance

While all the aforementioned fusion techniques have been applied in RS, there is no general consensus on a single technique that consistently delivers superior performance. However, most comparative studies indicate that intermediate fusion generally outperforms early and late fusion. Among the notable studies supporting this conclusion, [134] investigates two land cover and land use datasets, one combining hyperspectral data with LiDAR, and another using multispectral and SAR data, [97] combines satellite imagery form Sentinel-1 and Sentinel-2 missions for classification and segmentation tasks, [84] integrates optical and SAR data for deforestation detection, [283] combines arbitrary numbers of RGB aerial images for urban land use mapping, and [47] investigates two land cover and land use datasets, one including hyperspectral data with LiDAR, and another combining multispectral data and elevation maps.

2.2 EXPLAINABILITY FOR EARTH OBSERVATION

2.2.1 Terminology

Transparent AI

Interpretability & Explainability

Common usage

Within the domain of transparent AI , numerous terms and notions have been introduced in the last decade to define different aspects of solving algorithmic opacity. Among these, *interpretability* and *explainability* are frequently cited as key approaches to achieving model transparency. Table 2.2 collects certain definitions of the terms *transparency*, *interpretability*, and *explainability* from the literature, revealing a lack of consensus on their precise conceptual delineation [52, 285, 281]. In practice, however, *transparency* typically refers to the degree of clarity regarding the components and functioning of a model, whereas the terms *interpretability* and *explainability* are often used interchangeably [31, 215]. In this the-

sis, we adopt the definitions proposed by Palacio et al. [240], following their comprehensive review of a large collection of definitions from the literature:

Adopted definitions

> **Definition:** *An **explanation** is the process of describing one or more facts, such that it facilitates the understanding of aspects related to said facts (by a human consumer).*

> **Definition:** ***Interpretation** is the assignment of meaning (to an explanation).*

Accordingly, when a distinction is made in this thesis between both terms, *explanations* will refer to the output of the XAI method used (e.g. feature attributions or heatmaps), while *interpretation* will refer to the meaning and insights derived from analyzing the explanation output and their implications on understanding the model.

Another aspect of the XAI taxonomy is the common distinction between *black-box* and *white-box* models. *White-box models*, also known as *transparent* or *intrinsically interpretable* models, "provide their own explanations, which are faithful to what the model actually computes" [258]. These models are often associated with *ante-hoc explainability*, referring to the explanation readiness as a component of the model's design and nature. In contrast, *black-box models* are too complex for any human to comprehend. While the model architecture and learning process may be understood by the modeling engineer, the patterns and reasoning processes learned by the model remain opaque [258, 72]. In such cases, we often also talk about *post-hoc explainability*, referring to explainability methods applied to opaque models after their design [281]. Among post-hoc explanation methods, further distinctions are made between *model-agnostic* methods and *model-specific* methods. This difference is relative to the applicability of the methods, whether they can be applied to any ML model or are tailored to specific algorithm classes or architectures.

white-box models

Black-box models

Model-agnostic vs. model-specific

A further distinction in XAI methodologies concerns the explanatory scope, categorized as either *local* or *global*. Conventionally, *local explanations* analyze individual predictions to reveal the rationale behind specific model decisions, while *global explanations* seek to elucidate the complete reasoning framework of the model's inference process [318, 281, 72].

Local vs. global explanations

2.2.2 Explainability Objectives

Within the context of building transparent AI systems, XAI provides tools to achieve multiple objectives. Researchers and practitioners in the field have identified and described various goals for model explainability [123, 3, 68]. In this thesis, we focus on two primary objectives: *justification* and *improvement*.

XAI Objectives in literature

Justification consists of revealing the reasoning behind a decision made by the AI model [3]. This objective has been referred to as the "need for reasoning" by Hassija et al. [123] and "scientific understanding" by Doshi-Velez and Kim [68]. It primarily benefits end-users who seek to comprehend the rationale behind specific model outcomes. Justification methods are typical local explanation methods, explaining individual

XAI for justification

Tab. 2.2.: XAI nomenclature in the literature.

Transparency	"*clearly describing the model structure, equations, parameter values, and assumptions to enable interested parties to understand the model*"	[74]
	"*level to which a system provides information about its internal workings or structure*"	[304]
	"*the opposite of opacity or blackbox-ness. It connotes some sense of understanding the mechanism by which the model works.*"	[188]
	"*the processes that extract model parameters from training data and generate labels from testing data can be described and motivated by the approach designer*"	[255]
Interpretability	"*the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of.*"	[222]
	"*the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model*"	[227]
	aims to "*present some of the properties of an ML model in terms understandable to a human*"	[255]
Explainability	"*the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression).*"	[222]
	"*level to which a system can provide clarification for the cause of its decisions/outputs*"	[304]
	"*providing a way to improve the understanding of the user, whomever they may be.*"	[52]
	"*any information that can help the user understand and communicate why a model exhibits some pattern of decision-making and how individual decisions come about.*"	[251]

predictions. This is because end-users are generally more concerned with understanding specific cases they encounter rather than the broader mechanics of the model's reasoning. Popular techniques for providing justifications include feature attribution methods, which identify the input variables or regions of an input image that most influenced the model's prediction. Additional methods include counterfactual explanations, which indicate how an input could be modified to achieve a desired alternative outcome [224, 115].

Common XAI methods

Improvement represents a golden end-goal in AI and ML development [3]. This objective has also been referred to as the "need for advancement" in [123]. Through the justifications of individual predictions and the understanding of the general model reasoning, XAI implementation enables practitioners to identify weaknesses and biases in the model that may hinder optimal performance. Through the effective use of explanation results, practitioners can adjust both the data processing pipeline and the modeling stage, and subsequently, enhance the performance of the AI system. While various explanation techniques can be employed to achieve this objective, the data-centric paradigm in ML offers a particularly robust framework for incorporating explainability insights into the data preparation stage, to indirectly improve the system outcomes [202, 256].

XAI for improvement

Data-centric ML

Additional XAI objectives have been identified in the literature and achieved in real-word applications. These include the **discovery** of new laws in biology, physics, and chemistry [3, 123, 255], the **control** of model predictions to ensure safety and reasonable outcomes [3, 68], and the alignment with **fairness** and ethical guidelines of AI systems [123, 68]. All these objectives, including the justification and improvement, have been achieved through XAI for EO applications, as demonstrated by the examples in Table 2.3.

Other XAI objectives

Usage in EO

2.2.3 *Feature Attribution Methods*

A common approach of model explainability consists of assigning a score for each feature to quantify its relevance to the model [222]. These scores can also indicate the degree to which a feature supports or opposes a predicted label [234, 352], or measure a feature's saliency, defined as its capacity to cause a significant response or influence on the model's output [251]. Feature attributions are either *global* or *local*: when practitioners aim to understand the general relationships present in a dataset and how these are captured by a model, they rely on dataset-level interpretations that estimate global feature scores. These scores represent the overall contribution of individual features to the model's predictions across the dataset [227]. In practice, however, practitioners investigate local explanations, which focus on individual predictions and estimate feature importance scores specific to the corresponding input sample. Despite the methodological differences between global and local approaches, prediction-level (local) methods can be aggregated to derive dataset-level (global) insights [76, 30, 147]. In this thesis, the term *feature attribution* will refer to local explanations, unless otherwise indicated. Feature attribution methods can be further classified into two primary categories,

Definition

Global vs. local attributions

Mechanism types of attribution methods

Tab. 2.3.: Examples of RS studies achieving different objectives of XAI.

Justify	[39]	Feature attributions are estimated to explain a land use classification model based on satellite time series. The aim of this study is to make the model decisions auditable, and align with the Common Agricultural Policy (CAP).
	[161]	A human footprint index is estimated using a CNN model. Layer-wise Relevance Propagation (LRP) is subsequently applied to visualize the relevant features in the satellite input image, ensuring compliance with existing explainability regulations and verifying the model's alignment with domain knowledge.
	[145]	An example-based approach is implemented to explain a satellite image classification task and justify the model's predictions. Based on the examples returned by the explanation method, the end-user can assess whether the input lies within the manifold of the training data distribution or not.
Improve	[26]	Gradient-weighted CAM (Grad-CAM) is applied to explain a volcanic deformation detection model and identify the causes of false-positive predictions. The data is then augmented with targeted scenarios, resulting in an improvement in the model's accuracy.
	[163]	A satellite image classification model is explained by generating heatmaps at multiple intermediate layers and aggregating them. Inconsistent heatmaps across layers are interpreted as indicators of uncertain explanations. These uncertain results are reviewed and corrected by an expert supervisor, and the refined general heatmap is fed back into the model learning process to improve its performance.
	[33]	A gradient-based attribution method is applied to identify the most important features for estimating oceanic chlorophyll. Follow-up experiments demonstrated that using only the most sensitive bands as inputs to the model outperformed other baseline models.
Control	[259]	Saliency method is used to measure temporal importance in a crop classification task. Comparing results across different models facilitated an informed selection of the inference model based on its ability to rely exclusively on relevant time steps.
Discover	[75]	Sensitivity maps are generated to explain a classification model distinguishing protected areas from anthropogenic regions. The results are utilized to explore the characteristics of protected areas and extract scientific insights into defining the ambiguous concept of wilderness.

based on the mechanism by which the influence of a feature is estimated: *perturbation-based* and *backpropagation-based* methods.

Perturbation-based methods In order to estimate feature attributions in a model-agnostic manner, a common approach consists of perturbing the input features and assessing the impact of these changes on the model's prediction. Based on how the perturbation is applied, several methods have being proposed in the literature. *Permutation Importance* [35, 86] randomly shuffles the values of a feature across the dataset and measures the resulting impact on the model's outcome. The magnitude of prediction degradation provides an estimate of the feature's importance. The *Partial Dependence Plot (PDP)* [91, 124] visualizes the marginal effect of a feature on the model's predictions. By averaging predictions over all possible values of a feature, the method generates a plot to show the feature's influence. PDPs can also extend to combinations of two or three features, displayed using two- or three-dimensional plots with a color map to represent model predictions. Unlike PDPs, *Individual Conditional Expectation (ICE)* [112] provides local explanations by plotting predictions for a single feature value across all samples, with one curve representing each sample. This setup can however only handle a single feature per plot. Both PDP and ICE assume feature independence. To address this limitation, Accumulated Local Effects (ALE) [14] plots evaluate the effect of a feature within small intervals. The method calculates prediction differences between the interval boundaries and averages these differences across all data points in the interval. The prediction differences are displayed in a plot. ALE is particularly effective in handling feature dependencies and shows the magnitude and direction of a feature's effect. While the aforementioned methods are mainly suitable for tabular data, *Occlusion sensitivity* [340] is a perturbation-based method suitable for image data. It occludes specific regions of an image (e.g., replacing pixel values with gray or black) and measuring the impact on the model's predictions. By systematically occluding different regions, a heatmap can be generated to highlight important elements in the image. Extensions of this approach modify the occluded patch's size, shape, and sampling strategy [349, 181, 87, 353]. Another method similar to occlusion method, but usually applied on tabular data, is the *Shapley values* [271], derived from cooperative game theory. This local explanation method estimates the contribution of each feature to a prediction by computing the marginal contribution of the feature across all possible subsets of features. While theoretically sound, calculating exact Shapley values becomes computationally infeasible for datasets with many features due to the exponential growth in subsets. Sampling-based approximations are commonly employed to address this limitation [286]. Other solution derived from a similar concept are proposed in the literature [200, 199, 293, 130].

Permutation Importance

PDP

ICE plots

ALE plots

Occlusion sensitivity

Shapley values

A subcategory of perturbation-based methods, namely the **local approximation** methods, is highlighted in certain studies [251, 132]. Such techniques explain complex models by training simpler, interpretable models, such as linear regressors, and using their interpretation to provide insights into the larger model's behavior. A well-known variant

SHAP of Shapley values, namely the SHapley Additive exPlanations (SHAP) method [200], also referred to as KernelSHAP, decomposes a model's prediction into a weighted sum of the input features, where the weights represent feature attributions. SHAP relies on a perturbation-based approach to sample input instances and fits a linear model to estimate the contributions of individual features. Another method which decomposes model predictions into additive contributions from input features

LIME is Local Interpretable Model-agnostic Explanation (LIME) [253]. Unlike SHAP, LIME samples instances specifically from the local neighborhood of the input sample in the input space, weighting them based on their proximity to the input. LIME also offers implementation flexibility to practitioners, who can choose different types of interpretable models (e.g., decision trees or sparse linear models) to fit the local neighborhood. Practitioners can also customize the definition and sampling strategy of the local neighborhood.

Backpropagation-based methods Another family of model-specific methods leverages the internal structure and computation of a model to propagate a signal from the output back to the input and estimate feature

Saliency importance scores. *Saliency* [276] is the most basic form of such methods. It computes the gradients of the model's output with respect to its input. The resulting gradients indicate how sensitive the prediction is to changes in each input feature, highlighting both positive and negative contributions. For classification tasks, this involves evaluating the gra-

Input×Gradients dient of each predicted class probability. *Input × Gradients (I*G)* [275] improves upon basic saliency by incorporating the magnitude of the input features directly into the importance score. It multiplies the input features element-wise with their corresponding gradients. However, it shares some of the limitations of saliency, such as sensitivity to noise in the gradient computation [105] and the saturation problem [294].

Integrated gradients *Integrated Gradients (IG)* [294] addresses these limitations by incorporating multiple input values and their gradients. Instead of using a single gradient, IG accumulates gradients along a path from a user-defined baseline input (e.g., a zero or neutral input) to the input being explained. This approach further guarantees certain axioms and desirable properties in the estimated feature importance. Note that all gradient-based

LRP methods require that the model is differentiable. *LRP* [21] is another attribution methods based on signal backpropagation, which decomposes the output prediction into contributions of the individual input features without using gradients. It propagates the output backward through the network, adhering to specific conservation rules that ensure the sum of attributions equals the model output. These rules are adaptable to different neural network architectures, including convolutional, recurrent, and fully connected networks [176, 16, 17, 65, 221]. *Class Activation*

CAM *Mapping (CAM)* [348] is another method designed for CNNs with a specific architecture, where global average pooling is applied at the last convolutional layer, followed by a linear prediction head. It propagates the output signal backwards only until the last convolutional layer of the network, in which the activation maps are weighted and averaged to generate a heatmap, indicating the regions of the input image most re-

sponsible for the prediction. To mitigate the limitation of the architecture of the prediction head, *Grad-CAM* [267] extends CAM by relying on the gradient of the output with respect to the final convolutional layer, using them to weigh the activation maps and generate visual explanations. While originally designed for image classification, CAM and Grad-CAM can also be applied to regression tasks and tabular data.

Grad-CAM

2.2.4 Implementation in the Literature

Usage and Trends With the increasing adoption of ML and DL in RS, the use of complex models has become more common, driving a corresponding rise in the implementation of explainability techniques, as shown in Figure 2.1. Details of the search query used to generate this figure are described in Appendix A We conducted an in-depth review of 207 related papers, published in [132], and the outcome has revealed important trends of the usage of XAI in RS. We summarize the highlights in the following points:

Increasing usage of XAI

Systematic review

- **RS tasks:** Approximately 40% of the reviewed papers focus on three main applications: land cover mapping, agricultural monitoring, and natural hazard monitoring. This observation highlights the importance of model transparency in the particular context of land and agricultural activities.
- **XAI methods:** Among the various XAI techniques, SHAP emerges as the most commonly used method in RS research. We assume this is attributed to the method's accessible implementation through well-documented coding packages and its theoretical grounding in cooperative game theory, which makes it an easy off-the-shelf explanation option for most practitioners in this field.
- **ML models:** Due to the imagery nature of satellite data, most applications use convolutional networks, often paired with model-specific XAI methods such as CAM and Grad-CAM. Nevertheless, there are numerous studies which use tabular satellite-derived data, for which RF are mostly used, along with SHAP attribution method.

Feature Attribution for Earth Observation Numerous XAI methods providing feature attributions are originally designed for computer vision tasks: the features are the input image pixels, and the resulting attributions are visualized as heatmaps, overlaid on the input image. Consequently, it is common to refer to feature attribution techniques as visualization methods [251, 234, 4]. Feature attributions in computer vision allow practitioners in RS to easily examine image regions that significantly influence model predictions, enabling them to assess whether the model's reasoning is plausible or biased. Vasu and Savakis [315] applied CAM methods to explain three different CNN architectures on three benchmark datasets for land cover mapping, two of which included aerial images. Gawlikowski et al. [100] investigated the effects of haze and cloud coverage on a scene classification task. Using Grad-CAM, they explained misclassifications of a model trained on cloud-free satellite

Computer vision

Examples in RS

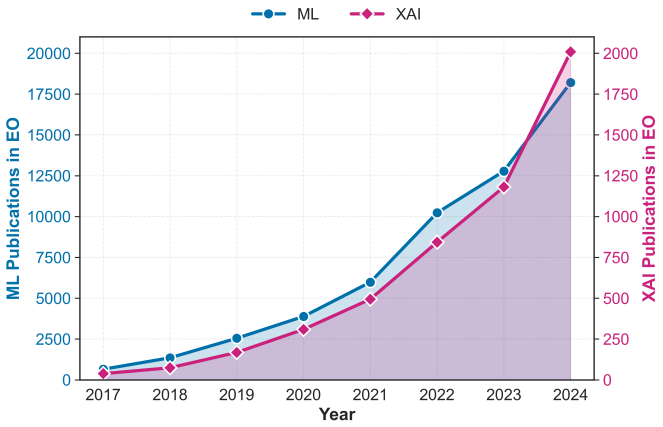


Fig. 2.1.: Number of publications of ML and XAI in EO.

images when applied to cloudy images. Oveis et al. [238] used LIME to assist human decision-making in a target recognition model based on SAR images. They also developed a metric derived from LIME explanations to measure the reliability of model predictions, and demonstrated its effectiveness in revealing the model's strengths and weaknesses. Ge et al. [102] adjusted and applied Grad-CAM across multiple layers of a convolution-based segmentation network to perform rice area mapping using SAR images.

Tabular data

While numerous applications in RS typically use satellite imagery as inputs, certain tasks transform the satellite data into tabular data. This transformation can occur either by processing images pixel-wise or by aggregating image data into single scalar values. Additionally, using time series of satellite data also diverges from the conventional usage of XAI in computer vision. Overall, the research focus on XAI for tabular data is relatively limited compared to imagery data [261]. Among RS applications that explain models trained on tabular data: Al-Najjar et al. [7] trained RF and SVM models for landslide prediction using time series of SAR features, vegetation indices derived from multispectral satellite data, and terrain information. They applied the SHAP method to evaluate feature importance, analyzing the agreements and differences in the results between RF and SVM. Aydin and Iban [19] similarly used RF and ensemble-based tree models for flood susceptibility prediction, and applied SHAP to identify the most important predictive features. Zhou et al. [351] predicted soil texture using time series of multispectral satellite images and terrain elevation data with an SVM model. Their models were applied to data points (i.e., pixels) rather than complete images, and SHAP was used to identify the features with the greatest influence on the model predictions. Andresini et al. [12] compared the performance of RFs, SVMs, and ensemble-based tree models for mapping insect infestations in European forests using multispectral satellite images processed pixel-wise. They used SHAP to explain each model, and compared the results across different models.

Examples in RS

Challenges and Limitations Several limitations in the use of XAI within RS applications have been identified in our systematic review [132]. For instance, while numerous studies employ ensemble or large tree-based models as inherently interpretable, the complexity and size of such models often surpasses the human capacity to directly investigate their inner-working. Consequently, these studies often fail to provide a detailed analysis of the interpretations generated by the models. In other cases, the selection of XAI methods is often neither justified nor motivated. Despite the availability of numerous explanation techniques in the literature, their results do not necessarily align, highlighting the importance of a reasoned selection process or prior evaluation of different methods for the specific task at hand. Furthermore, in some cases, it is assumed that the explanations imply causal relationships between inputs and targets. For instance, when a variable in tabular data is assigned a high importance score, researchers may conclude that the variable is a significant causal factor for the predicted event. However, identifying cause-and-effect relationships is beyond the scope and capability of most XAI techniques, which are designed to evaluate the importance of variables for the model's predictions rather than for the underlying phenomena being studied. A related misleading interpretation of the explanation results is the reliance on anecdotal evidence; a common evaluation strategy involves selecting (or cherry-picking) individual examples that appear plausible to practitioners and align with their intuition. However, many researchers argue that relying on unverified intuition and anecdotal inspection is not a robust method for validating model explanations [4, 67, 232, 146, 10, 336]. Instead, quantitative evaluation methods are regarded as more reliable for ensuring the correctness and robustness of explanations.

Large interpretable models

Method selection

Causal interpretations

Anecdotal evidence

2.3 SUMMARY

In this chapter, we introduced key concepts that are important for the understanding of this thesis.

AI in EO Given the inherently data-rich nature of EO, Section 2.1 examines AI's role in this domain while tracing the historical evolution of AI networks - from traditional ML to modern DL approaches - within EO applications. This section further discusses the crucial role of multimodal learning for processing diverse RS data modalities.

Explainability in EO In Section 2.2, we highlight the importance of XAI in EO research, reviewing both current applications and identified limitations from existing literature.

This summary concludes the background of this thesis. Further background information and related work are reported at within each chapter for the specific topic that we consider.

PartII

XAI FOR JUSTIFICATION

3 Multimodal Learning

3.1 INTRODUCTION

The diversity and accessibility of data in RS have created opportunities to apply multimodal learning techniques, enhancing model performance across various EO tasks. However, not all tasks demand high levels of complexity. For example, in agronomy, some traits can be accurately predicted from a single sensor and using basic modeling systems. As highlighted by Weiss et al. [326], agronomic traits such as typological, physical, chemical, biological, structural, or geometrical properties can be inferred from remote sensors at varying levels of complexity, as illustrated in the graph reproduced in Figure 3.1.

Agronomic traits inferred from remote sensors

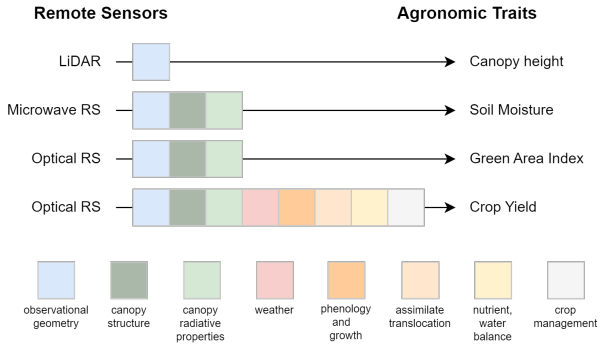


Fig. 3.1.: Levels of complexity involved in predicting certain agronomic traits from remote sensors observations. The driving factors connecting the data to the target values are represented in colored boxes. Figure reproduced from [326].

Yield prediction is a complex task influenced by numerous factors and interrelated biophysical processes, as shown in Figure 3.1. To address this challenge, we design in this chapter multimodal networks that incorporate multiple data sources for yield prediction, with a particular focus on explaining the model's behavior. Specifically, we employ XAI techniques to **justify** the model's decisions by analyzing how input features and modalities contribute to the final predictions.

Yield prediction task

In Section 3.2, we begin with a straightforward modeling framework where input modalities are aligned during pre-processing and treated as a single modality. This setup aligns with standard architectures for which XAI methods are typically designed, allowing us to evaluate and compare various post-hoc feature attribution techniques to reveal the model's internal reasoning. In Section 3.3, we explore more advanced modeling architectures employing intermediate fusion techniques, and

Chapter structure

we leverage the components of Transformer-based architectures to intrinsically explain the model.

3.2 POST-HOC EXPLAINABILITY WITH EARLY MODALITY FUSION

Ease of explainability under the early fusion setup

When using multiple modalities as inputs to achieve a specific task, conducting an early fusion (i.e. input fusion) of the modalities significantly facilitates the modeling task, given that any network architecture designed for a single modality becomes applicable. Consequently, explanation methods originally designed and tested on single-modal networks can readily be applied as well. Hence, we begin our experiments regarding the explanation of agricultural tasks under the early fusion setup. Specifically, we apply and evaluate multiple feature attribution methods, and leverage the results to explain the challenging task of yield prediction. Given the high resolution of the input data used, we predict the yield at the pixel-level. Subsequently, we only consider attribution methods applicable to multivariate time series.

3.2.1 Attribution Methods

Multiple feature attribution methods were presented in the previous chapter, Section 2.2.3. We select a representative subset for implementation in this chapter and use Captum package [165] within the PyTorch modeling framework [13]. Implementation details are described in the following.

Let X and $\bar{X}$ be two input samples, each compromising a single or multiple modalities fused together. We consider that $\bar{X}$ is a *baseline* sample, whose significance and selection criteria we discuss subsequently. Let i be an input feature, and x_i the value of X at this feature. Given a deep network f, let $\hat{y} = f(X)$ be the prediction returned by the model, given X. The goal of the attribution method is to determine the contribution $a_i \in \mathbb{R}$ of each input feature i of X to the output $\hat{y}$. In the following, we briefly describe the methods used to estimate the contributions a_i.

- **Occlusion** [340] replaces the value of feature x_i by its baseline value $\bar{x}_i$ and compute the difference in output, i.e. $a_i = f(X) - f(X_{|_{x_i=\bar{x}_i}})$.
- **Saliency** [276] returns the gradient of the output with respect to the input feature x_i: $a_i = \frac{\partial \hat{y}}{\partial x_i}$.
- **I*G** [275] multiplies the input feature value by the gradient of the output with respect to that feature: $a_i = x_i \cdot \frac{\partial \hat{y}}{\partial x_i}$.
- **IG** [294] approximates the integral of gradients of the model's output $f(\bar{X} + \alpha(X - \bar{X}))$ with respect to the input X along the path (straight line) from $\bar{X}$ to X, i.e. $\alpha \in [0, 1]$. The Gauss-Legendre quadrature rule [113] is used to approximate the integral, using 100 steps, and the outcome is multiplied by the input $(X - \bar{X})$ [11].
- **GradientShap** [200] is a gradient method to approximate Shapley values. It adds Gaussian noise, $\mathcal{N}(0, 1)$, to each input sample multiple times, selects a random point along the path between $\bar{X}$ and X, and computes the gradient of outputs with respect to those selected random points. The expected (i.e. average) value of 25 gradients is multiplied by $(X - \bar{X})$.

- **Shapley Value Sampling (SVS)** [286] involves taking each permutation of the input features and adding them one-by-one to the baseline. The output difference after adding each feature corresponds to its contribution, and these differences are averaged over all permutations to obtain the attribution. This computationally intensive process is approximated by sampling 50 random permutations and averaging the marginal contribution of features based on these permutations.

- **LIME** [253] trains an interpretable surrogate model by sampling 50 points around X in the input space and using model evaluations at these points to train a simpler surrogate model. The method returns coefficients of the linear model as the attribution values. We use a linear model trained with L1 prior as regularizer (i.e. the Lasso), setting $\alpha = 0.001$ to control regularization strength. Input samples around X are created by adding a Gaussian noise $\mathcal{N}(0, 0.01)$ to X (at non-padded instances). In the training of the surrogate model, the samples are weighted according to their similarity to the input X, which we estimate using an exponential kernel based on L2 norm, with a kernel width of 1.1.

- **KernelShap** uses LIME framework to approximate Shapley values, as described in [200, 251]. We set the number of samples of the original model used to train the surrogate interpretable model to 50.

What is a baseline vector?

Attribution baselines Most of the attribution methods require a user-defined baseline vector $\bar{X}$, which can serve different purposes: it is used by LIME, Occlusion, SVS and KernelShap to replace occluded features at computation time, and it serves as a starting point from which the integral in IG and the gradients expectation in GradientShap are computed. In this work, we introduce and compare two distinct baselines for the attribution methods, which we term the *padded baseline* and the *mean baseline*.

Padded baseline

The padded baseline essentially entails a vector that is fully padded with the -1 value. Such baseline vector is the same for all inputs X. Note that we generally use the value -1 to represent the absence of data during periods outside the crop growth season, or at months between the seeding and harvesting dates where no satellite data is available (e.g. because of the clouds).

Mean baseline

The mean baseline is computed as the mean vector across the entirety of the dataset. We further adjust this vector to each input X, by matching the missing time steps padded with -1 in both X and $\bar{X}$, in order to account for the varying presence of missing satellite data at some early and/or late time steps within each field. In case this adjustment is not applied, we would expect scenarios where the occluded value stems from a missing time step, but is substituted with a high value taken from the mean vector across the whole dataset. As a result, a substantial value shift occurs. This can potentially distort predictions in a misleading manner, implying a high importance of the occluded feature. By adjusting the mean baseline vector, any absent feature maintains a consistent padding

Why adjust the baseline?

value when subjected to occlusion, resulting in its importance remaining effectively negligible.

Attribution evaluation We use two evaluation metrics to assess the attribution results and compare different methods: the *sensitivity* and *infidelity* metrics [336]. Each score assigns a single value for each data sample to assess the quality of the computed attribution.

Infidelity score

More specifically, given $X \in R^N$, an attribution function $\Phi : \mathcal{F} \times \mathbb{R}^N \rightarrow \mathbb{R}^N$, and a meaningful perturbation $I \in \mathbb{R}^N$ with a probability measure μ, the infidelity of the attribution function Φ for input sample X and prediction function f is defined as:

$$\text{INFD}_\mu(\Phi, f, X) = \mathbb{E}_{I \sim \mu}\left[\left(I^T \Phi(f, X) - (f(X) - f(X - I))\right)^2\right]$$

Sensitivity score

Sensitivity measures the extent of the attribution change when the input is slightly perturbed. It is measured using Monte-Carlo approximation by sampling points within an L_p-ball neighborhood of X with radius $r > 0$. Formally, for a given perturbation radius $r \in \mathbb{R}^+$, the sensitivity metric is defined as the maximal normalized difference in attributions:

$$\text{SENS}_r(\Phi, f, X) = \max_{\|X' - X\| \leq r} \frac{\|\Phi(f, X') - \Phi(f, X)\|}{\|\Phi(f, X)\|}$$

Scores interpretation

Overall, the two methods quantify the instability of the attribution of an input X by perturbing this input into X', computing the attribution of X', and comparing both attributions. The sensitivity score evaluates the degree to which the explanation is affected when the input is slightly perturbed, while the infidelity score computes the expected difference between (i) the dot product of the input perturbation to the explanation vector $\{a_i\}$, and (ii) the outcome difference between X and X'.

Temporal & Spectral attributions Since our processing operates pixel-wise and incorporates temporal modalities, each input sample contains two key dimensions: spectral (bands) and temporal (time steps). Let B denote the number of spectral bands and T the number of time steps. For any given band $b \in \{1, \ldots, B\}$ and time step $t \in \{1, \ldots, T\}$, the attribution value $a_{b,t} \in \mathbb{R}$ quantifies feature $x_{b,t}$'s contribution to the prediction $\hat{y}$. The magnitude and sign of $a_{b,t}$ reflect how strongly and in what direction (increasing or decreasing) the feature influences the prediction. Importantly, these attributions relate specifically to the model's output rather than its loss function - thus, while positive attributions indicate features that increase $\hat{y}$, this does not necessarily correlate with improved model performance. In this chapter, all attribution results systematically go through two types of transformations before any further analysis, unless otherwise is specified. In the first transformation, denoted $\breve{a}_{b,t}$, we normalize the attributions for each pixel such that the sum of absolute values across all features equals 1. The second transformation, $\hat{a}_{b,t}$, which we apply in all the remaining analyses, uses the absolute values from the first operation. This final representation captures the features' absolute importance to the model's prediction - precisely the focus of our investigation:

Attributions processing

$$\check{a}_{b,t} = \frac{a_{b,t}}{\sum_b \sum_t |a_{b,t}|}, \quad \hat{a}_{b,t} = |\check{a}_{b,t}| \,. \tag{3.1}$$

Spectral & temporal aggregates

In addition, we can evaluate the spectral attribution $\mathrm{SA}(X)_b$ of band b and the temporal attribution $\mathrm{TA}(X)_t$ of time step t by summing up the transformed attributions along the temporal or spectral dimension:

$$\mathrm{SA}(X)_b = \sum_{t=1}^{T} \hat{a}_{b,t} \,, \quad \mathrm{TA}(X)_t = \sum_{b=1}^{B} \hat{a}_{b,t}. \tag{3.2}$$

Furthermore, the importance of a specific band or time step can be evaluated on the field level by averaging its attribution scores over all the pixels from that field, and similarly at the dataset level.

3.2.2 Dataset

In this section, we present the datasets used for yield modeling, describing the corresponding study sites, the processing of the target yield values and the characteristics of the input modalities used.

Study Sites

Crop farms location

We collected agricultural data from different regions in South America and Europe, as depicted in Figure 3.2. Specifically, we obtained soybean crop data from two countries: Argentina and Uruguay, involving 10 and 8 different farms, respectively. In Northern Germany, we collected data of wheat and rapeseed crop fields, all originating from the same 6 farms. In Argentina, all soybean farms used in our study are situated in the northern regions, close to Uruguay, while in Germany, rapeseed and wheat crops are cultivated in the same farms, located in the northwest and close to the center of the country.

Climate & topographic properties

To gain a comprehensive understanding of the climate conditions and topography in these regions, we rely on the Global Agro-Ecological Zones (GAEZ) model documentation, specifically for insights into land, water [1] and agro-climatic [2] resources [85]. All our study regions share a sub-tropic climate with moderately cool temperatures. The air is humid in North Argentina and Uruguay, sub-humid without significant soil or terrain constraints in German fields near the country's center, and semi-arid with notable soil and terrain limitations in northern Germany. The total number of rain days (days with daily precipitation exceeding 1 mm) varies across these regions. In North Argentina and Uruguay, it ranges from 100 to 130 rain days, with annual precipitations averaging around 1900mm and 1000-1900mm, respectively. In Northern Germany, the number of rain days varies between 160 and 220, with annual precipitations ranging between 500 and 700mm. The length of the growing season is approximately 200 days in Germany and extends to 300 days in North Argentina and Uruguay. Note that the annual precipitation values and growing period lengths are averaged over the period 1981-2010. As for the topography of the study sites, the median altitude exhibits significant

[1] https://gaez-services.fao.org/apps/theme-1/.
[2] https://gaez-services.fao.org/apps/theme-2/.

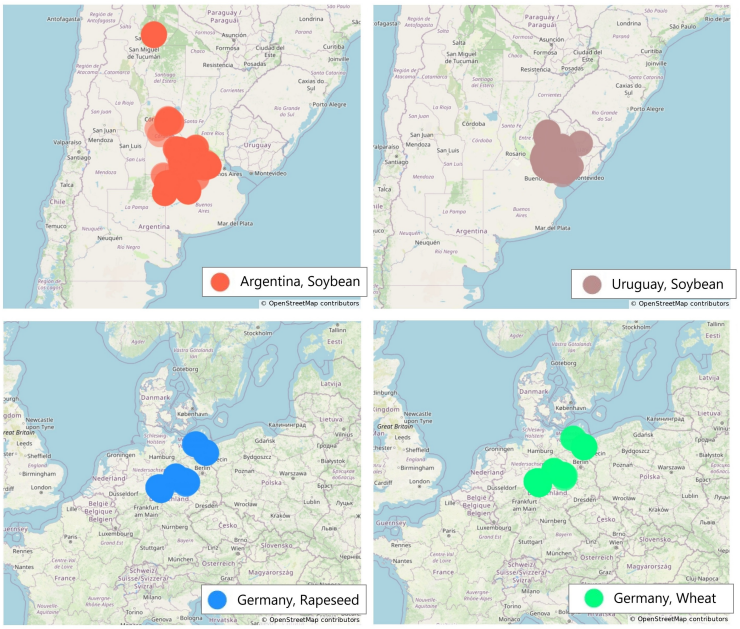


Fig. 3.2.: Geolocations of crop fields in South America and Europe: Argentina (soybean), Uruguay (soybean) and Germany (wheat and rapeseed). The buffer sizes are large to preserve data confidentiality. Map generated using OpenStreetMap data [236] and Plotly python package.

variations. In German fields, altitudes range from 40m in the northwest to around 250m near the country's center. In Uruguay, altitudes span from 30m to 200m, while in Argentina, the altitude variation is relatively smaller, ranging from 30m to 120m.

Yield Data

Yield datasets

In our study, we train ML models to predict crop yield using rasterized yield maps at 10-meter resolution. These data are organized into four datasets, as outlined in Table 3.1, with individual fields potentially represented across multiple years when yield data were collected in different seasons. Uruguay contributes the largest number of soybean fields, followed by Argentina and Germany. The variation in typical field sizes across countries accounts for the disparity between field counts and pixel numbers.

Yield data collection

The yield data is collected by combine harvesters equipped with yield monitors, which record georeferenced yield measurements (in t/ha) at high spatial resolution along harvesting paths. We processed these data by: (1) rasterizing yield points through averaging within 10×10m grid cells matching satellite image resolution, (2) applying crop-specific yield thresholds (15 t/ha for soybean, 10 t/ha for rapeseed, 20 t/ha for wheat) based on agronomic expertise, and (3) removing outliers via the three-sigma rule (excluding values beyond ± 3 standard deviations from the mean).

Tab. 3.1.: Overview of the crop yield datasets.

Dataset	Region	Crop Type	Years	# Farms	# Fields	# Pixels
ARG-S	Argentina	soybean	2017 – 2022	8	190	1.45M
URG-S	Uruguay	soybean	2018 – 2022	10	572	1.79M
GER-R	Germany	rapeseed	2016 – 2022	6	111	0.30M
GER-W	Germany	wheat	2016 – 2022	6	188	0.31M

Input Modalities

Satellite data

We supply the model mainly with multispectral satellite data, from the Sentinel-2 (S2) mission, and conduct additional experiments adding more modalities to the network. More specifically, we use the entire spectral signal of S2 Level-2A data and collect all scenes from seeding to harvesting. The additional scene classification layer (SCL) layer is used to identify and omit clouded pixels. Spectral bands with lower resolutions are upsampled to 10m resolution.

Additional modalities

Within the spatial boundaries of each field, we further collect weather data derived from the ECMWF Reanalysis (ERA5) [129], soil data from SoilGrids in 250m resolution, and digital elevation maps (DEMs) data from NASA's Shuttle Radar Topography Mission (SRTM) [82] in 30m resolution. Weather data is aggregated for each day at field level for minimum, maximum, and mean temperature and total precipitations. We use eight soil properties, i.e. cation exchange capacity (cec), volumetric fraction of course fragments (cfvo), nitrogen, soil pH (phh2o), sand, silt, soil organic carbon (soc), and clay, at depth of 0-5, 5-15, and 15-30 cm. For DEM, in addition to the elevation values, we derived the aspect, curvature, slope and the Topographic Wetness Index (twi).

All data were resampled to 10m resolution via cubic spline interpolation. Table 3.2 summarizes the complete feature set.

3.2.3 *Yield Modeling*

Modeling setup

Time Series Modeling To process the data pixel-wise, we create a multivariate time series for each pixel of the rasterized data maps. For the temporal sampling, we compare two approaches: *raw time-series* and *monthly sampling*. In the former approach, all the satellite data available between seeding and harvesting dates is used. Across the different datasets, the length of the growth period and the number of available satellite images depend on the crop type and the latitude coordinate of the field, which results in different revisit times: countries closer to the equator have a longer gap between two consecutive satellite images compared to farther countries. Thus, we fix the sequence length to 150 across all datasets and pad the initial time steps in case of shorter sequences.

Full raw time-series

Monthly sampling of the time-series

In the monthly sampling, the sequence length is fixed to 24 months, corresponding to two calendar years, such that the harvesting date always falls in the second year. This strategy is implemented to maintain unique indices for each month across various fields and years. However, only data from seeding to harvesting are collected, while the remaining time steps are padded, as described in [243]. In general, using a uniform

Tab. 3.2.: Characteristics of all input modalities used for crop yield prediction, with corresponding temporal and spatial resolutions.

Modality	Dynamic features	Spatial Res.	Temporal Res.
Satellite (S2)	B01 - Coastal Aerosol	60 m	5 days
	B02 - Blue	10 m	
	B03 - Green	10 m	
	B04 - Red	10 m	
	B05 - Red Edge 1	20 m	
	B06 - Red Edge 2	20 m	
	B07 - Red Edge 3	20 m	
	B08 - NIR	10 m	
	B8A - Narrow NIR	20 m	
	B09 - Water vapour	60 m	
	B11 - SWIR 1	20 m	
	B12 - SWIR 2	20 m	
Weather (**ERA5**)	Max temperature	30 km	Daily
	Mean temperature		
	Min temperature		
	Total precipitation		

Modality	Static features	Spatial Res.
DEM (**SRTM**)	Elevation	30 m
	Slope	
	Curvature	
	TWI	
	Aspect	
Soil (**SoilGrids**)	CEC	250 m
	CFVO	
	Nitrogen	
	pH-H2O	
	Sand	
	Silt	
	SOC	
	Clay	

sequence length further facilitates the processing of the extensive dataset in mini-batches during model training and validation. When training with additional modalities under the monthly sampling scenario, static data (i.e. soil and DEM) are duplicated at each time step, while weather data is monthly aggregated before being stacked to the satellite time series. The name of each weather variable in Table 3.2 indicates the type of aggregation it received.

Temporal processing of additional modalities

LSTM-based Modeling In order to capture the temporal dependencies in the data, we use long short-term memory neural network (LSTM) networks [131] and train a separate model for each dataset. In each model, two LSTM layers are stacked on top of each other, with 128 hidden units each, and a dropout of 0.3% to prevent overfitting. The output of the last time step of the second LSTM layer is fed into a sequence of operations: The first linear layer has 128 neurons, followed by batch normalization and a ReLU [229] activation function. The second linear layer with 128 input features and a single output feature is then applied to make the final yield prediction. Model optimisation is based on the mean absolute error (MAE) loss.

Training For each dataset, we train three models, each under a different temporal sampling strategy: raw time-series (*raw-ts*), monthly sampling (*24ts*), and monthly sampling with additional data sources (*24ts+ads*). The training is conducted under the cross-validation setting using 10 folds, grouped by fields and stratified by farms, to ensures that pixels from the same field are always grouped in the same fold. To evaluate the quality of the prediction and the model performance, we use MAE, root mean square error (RMSE), and the coefficient of determination (R^2) metrics described in Appendix B. In the results, we report validation metrics averaged across the 10 folds.

Model evaluation

We first evaluate the performance of the yield prediction models, where one model is trained for each dataset and for each sampling and feature composition setting. The results, based on the evaluation metrics used, are summarized in Tables 3.3 and 3.4 for the raw and monthly samplings, respectively.

Field- and subfield-level evaluation

We conduct the evaluation at both the subfield and field levels. For the former, we compute each metric for each pixel and then average over the entire dataset. In the latter case, we first average the predicted and target yield values over the pixels within each field, before computing the metrics for each field, and finally averaging over the entire validation set. Since we train the model using the cross-validation setup, the metrics reported in Tables 3.3 and 3.4 are the average across the 10 validation folds. In all our experiments, we consistently observe that the results at the field level outperform those at the subfield level. This disparity arises because subfield metrics measure the model's ability to capture the in-field variability of yield values. In contrast, on the field level, we

Tab. 3.3.: Modeling results under the raw time series on all datasets and with two different training data: satellite bands (S2) or vegetation indices (VI). The best scores in each dataset are highlighted.

Dataset	Bands	Field			Subfield		
		MAE	R2	RMSE	MAE	R2	RMSE
ARG-S	VI	0.41	0.72	0.55	0.67	0.62	0.90
	S2	**0.38**	**0.76**	**0.52**	**0.66**	**0.64**	**0.88**
URG-S	VI	**0.36**	**0.77**	**0.51**	**0.78**	**0.41**	**1.22**
	S2	**0.36**	0.75	0.53	**0.78**	**0.41**	**1.22**
GER-R	VI	0.71	0.53	0.92	1.04	0.31	1.38
	S2	**0.61**	**0.67**	**0.77**	**0.98**	**0.38**	**1.31**
GER-W	VI	1.45	0.11	1.79	2.21	0.07	2.84
	S2	**0.91**	**0.55**	**1.27**	**1.73**	**0.35**	**2.38**

Tab. 3.4.: Modeling results under the monthly temporal sampling on all datasets and with three different training data: satellite bands (S2), satellite bands and additional modalities (S2+ADS) and vegetation indices (VI). The best scores in each dataset are highlighted.

Dataset	Bands	Field			Subfield		
		MAE	R2	RMSE	MAE	R2	RMSE
ARG-S	VI	0.46	0.64	0.63	0.74	0.55	0.98
	S2	0.4	0.74	0.53	0.69	0.61	0.92
	S2+ADS	**0.38**	**0.76**	**0.51**	**0.66**	**0.63**	**0.89**
URG-S	VI	0.44	0.64	0.64	0.82	0.36	1.27
	S2	0.4	0.69	0.59	0.8	0.38	1.25
	S2+ADS	**0.35**	**0.76**	**0.52**	**0.78**	**0.41**	**1.23**
GER-R	VI	0.62	0.64	0.81	1.02	0.34	1.36
	S2	0.6	0.65	0.8	1.01	0.35	1.35
	S2+ADS	**0.55**	**0.69**	**0.75**	**0.96**	**0.41**	**1.29**
GER-W	VI	1.31	0.23	1.66	2.07	0.17	2.68
	S2	**0.91**	0.57	1.25	1.79	0.31	2.45
	S2+ADS	0.92	**0.59**	**1.22**	**1.74**	**0.33**	**2.41**

notice that averaging the individual predictions can more easily and accurately estimate the average yield per field.

Comparing performance across datasets

When comparing the R^2 results across the different datasets, we note that the top scores at the field level are generally comparable across the datasets, except for GER-W, which consistently ranks last by a considerable margin. At the subfield level, the results achieved in ARG-S clearly outperform those in other datasets, regardless of the temporal samplings used, while GER-W is still not reaching comparable scores. These disparities may originate from the distinct agronomic characteristics of each crop. Therefore, it is possible that the input data in our experiments might be inherently insufficient to capture these unique characteristics and establish a strong connection with the target yield, resulting in less accurate predictions. Furthermore, satellite data availability can also vary from one country to another, influenced by factors such as geolocation and cloud coverage related to the climate in each region [287]. The quality of the yield data collected from different countries may also further impact the overall results, given that the data providers vary from a country to another.

Impact of adding modalities

Analyzing the impact of enriching the satellite data with additional modalities, as shown in Table 3.4, reveals an overall improvement in the results across all datasets. This results also align with comparable studies [211, 217, 213, 243]. The smallest increase is observed in the Argentina

dataset, while the most significant impact is seen in the rapeseed dataset in Germany.

Comparing raw vs. monthly sampling

Regarding the influence of using the complete satellite time series compared to monthly selection, performance generally improves in most cases when using the same input modalities. However, the performance difference is not substantial, suggesting that both sampling methods capture sufficient patterns from the satellite data to predict crop yield. Still, raw sampling likely captures a few additional patterns: growth stages are better represented in the raw sampling, whereas the limited number of satellite instances in monthly sampling may under-represent certain stages. Moreover, the magnitude of improvement varies across datasets, making it difficult to draw a general conclusion applicable to different environmental contexts. For example, the smallest performance difference is observed in the German fields, likely due to high cloud coverage in satellite images caused by Germany's rainy weather.

3.2.4 *Interpretable Features for Interpretable Models*

Crafting meaningful features

As discussed by Guidotti et al. [116], the process of explaining a model through feature attribution methods requires a solid comprehension of the predictive features being employed. Conventional methodologies for crop modeling rely on the crafting of features by experts, with the aim of encapsulating the extensive array of variables recognized as direct yield drivers. However, when ML and DL models are used, the scenario changes. Raw information, acknowledged for its potential to serve as a proxy for crop growth and health, is directly supplied to the network with minimal to no alterations. In our work, we address this limitation of explaining ML models in two dimensions: inclusion of vegetation indices (VIs) and aggregation of the temporal dimension.

Vegetation Indices

Substituting raw bands with VIs

With the aim of augmenting the interpretability of the spectral dimension, we include a set of well-established VIs, frequently used in the literature for crop classification and yield prediction. Indices merge two or more spectral bands and generate an index with agronomic relevance, rendering it more easily interpretable than raw satellite bands. The main idea is to substitute the original S2 raw bands with pre-calculated indices during the training phase. Subsequently, we evaluate the impact of this adjustment on the model's performance, and compare it to training the model using raw inputs, before concluding on the efficiency of this method. The full list of indices used and their respective formulas are summarized in Appendix C.1.

Results & discussion

The results are included in Tables 3.4 and 3.3. When replacing the full spectrum of satellite data with VIs, we observe varying responses. In the monthly sampling, this change had a minor negative impact in URG-S and GER-R, but in contrast, it led to a substantial decrease in scores in ARG-S and GER-W datasets. Notably, in these two datasets, the R^2 score dropped by 10 and 34 percentage points (p.p), respectively, at the field level. When applying this same operation to raw time series data, a similar decrease in scores is observed in Germany's datasets, with

a 14 p.p drop in rapeseed fields and a 44 p.p drop in wheat fields at the field level. However, the use of VIs yielded comparable results to the full satellite spectrum in Argentina and even slightly improved results at the field level in Uruguay. One potential explanation for these results might be the need for complex spectral band interactions. Indices usually map only linear or low-complexity inter-band interactions, while neural networks are able to learn non-linear and more complex interactions.

Given the overall relatively poor performance achieved by indices-based models, such models will not be addressed in the explanation analysis in following sections.

Growth Stages

Aggregating attributions over growth stages

The second dimension we address to improve features' interpretability is the temporal dimension. To gain insights and derive explanations from crop yield models using time series, we can use different types of domain knowledge, with particular potential in the different phases that each crop goes through. Hence, we exploit the principal growth stages of the BBCH scale for rapeseed and wheat [173], and a modified scale for soybean [209]. We describe these scales in Appendix C.2. Approximations of the growth stage periods of each field were used[3]. Concretely, we first get the modified attributions $\hat{a}_{b,t}$ for each explained pixel X at each band b and time step t. We further define for each growth stage t_g a set T_g of the time steps it covers, then compute the total attribution as follows:

$$\hat{a}_{b,t_g} = \sum_{t \in T_g} \hat{a}_{b,t}.$$

Difference between spectral and temporal interpretability

Note that improving the temporal interpretability is applied post model training on the attribution score, in contrast to the spectral interpretability, which can only be applied on the input data, and thus requires training a separate model. Therefore, the analysis based on growth stages will be included in subsequent explanation analysis. A major benefit of this approach is that it overcomes the misalignment of the seeding and harvesting months between the different fields in each dataset, which will facilitate the comparison of temporal attribution results across different fields.

3.2.5 *Model Explainability*

Evaluation of Attribution Methods

Experimental setup

The soybean dataset from Argentina is used for the evaluation of multiple feature attribution methods, and the data is pre-processed according to the monthly sampling. To use explainability for scientific insights, higher model accuracies yield more informative and stable models. Hence, we specifically interpret the fold with the highest R^2 score, and compute the attribution of all the pixels in the fold-specific validation fields.

[3] Phenology data was provided by www.xarvio.com, using their in-house developed and commercially deployed growth stage models.

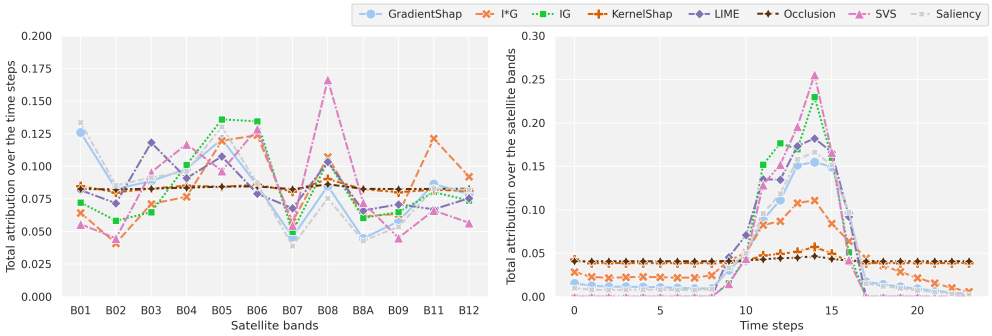


Fig. 3.3.: Feature attribution methods comparison results using the mean baseline and ARG-S dataset. On the left is the spectral importance, and on the right is the temporal importance.

Assessing the agreement of attribution methods

Attribution scores Using the mean baseline vector, we compute both spectral and temporal importances, as illustrated in Figure 3.3. We observe a consistent pattern in the temporal attributions. Specifically, time steps before seeding and after harvesting months (approximately from time step 9 to 16) exhibited low importances, while the first growing months had a more significant influence, and the late stages were the most important. However, looking at the spectral importance, it is hard to decipher an agreement pattern across the feature attribution methods. This highlights the need for further evaluation tools to assess these methods better.

Sensitivity and Infidelity scores

Attributions stability In Figure 3.4, we assess the stability of the methods and their robustness against both minor and significant input modifications. This is quantified using the infidelity and sensitivity scores. The boxplots show the distributions of both evaluation metrics over the attributions of all explained samples, when different attribution methods are used. Our goal is to identify a method for which the attributions yield low sensitivity and infidelity. The results reveal that most methods exhibit high sensitivity to minor input perturbations, with only four methods maintaining flat scores when subjected to significant noise in the input data, i.e. flat infidelity scores. Notably, SVS stands out as the sole method that remains robust against both perturbation types.

Qualitative assessment of the attribution stability

We further illustrate this characteristic in Figure 3.5 by examining the attribution maps (i.e. field-wide attribution scores) of a single feature for various methods. On the left column, the input values of the satellite band B08 from March of the second year are visualized on the first map, along with the reference and predicted yield maps of a field from ARG-S dataset. The remaining columns contain the attribution maps of the same feature and field when different methods are used. The blue and red values distinguish between positive and negative attributions, which have contributed to either increasing or decreasing the predicted yield value for each pixel, respectively. We observe that only SVS produces a smooth attribution map, in contrast to the other attribution techniques where the results show pronounced spatial fluctuations. We verified that this pattern is consistent across the remaining features. Such significant

SVS outperforms other methods

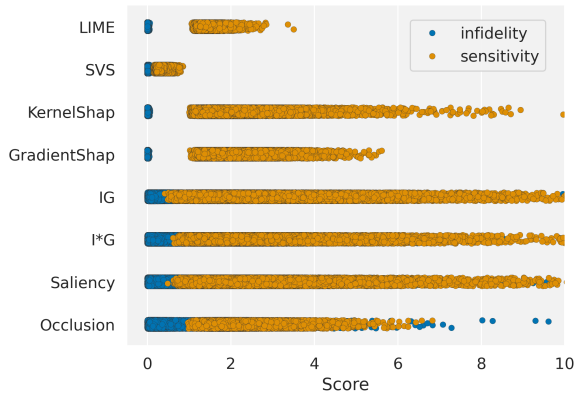


Fig. 3.4.: Quantitative evaluation of the feature attribution methods using the mean baseline on ARG-S dataset: Infidelity and sensitivity scores.

fluctuations, indicative of high sensitivity, is undesirable because adjacent pixels often share similar input features and target values, implying that their attributions should also exhibit similarity [10, 29].

Baseline comparison We further evaluate the choice of the baseline. While the results shown in Figures 3.4 and 3.3 remain similar when using the padded baseline vector (see Appendix D.1), examining the attribution maps highlights a fundamental difference between the two baseline types. In Figure 3.6, we select a field from the ARG-S dataset and visualize its attribution map for three satellite bands over three consecutive months. In Figure 3.6.a, the results using the mean baseline effectively differentiate between high and low yield regions by displaying positive and negative attribution values, matching the variance on the target yield map (refer to Figure 3.5). However, in Figure 3.6.b, the results of the padded baseline fail to capture similar patterns, with the maps consistently attributing values as either strictly positive or negative.

Comparing attribution maps of mean and padded baselines

Selected method We would like to finally highlight an additional important advantage of using SVS, which stems from game theory; Shapley values inherently accounts for feature interactions by approximating the impact of a feature when added to all possible subsets of the remaining features [199]. Based on our analysis of the evaluation results and the above discussion, we conclude that SVS is a reliable interpretation method for feature importance estimation, and we therefore conduct the remaining explanation analysis based on the SVS attributions and using the mean baseline.

Selection of SVS method and mean baseline

Spectral and Temporal Attributions

In this section, we investigate the interpretation results of the models trained on each dataset and closely examine the spectral and temporal importance of each model.

Experimental setup Given that SVS is a perturbation-based attribution

Selection of samples and model to be explained

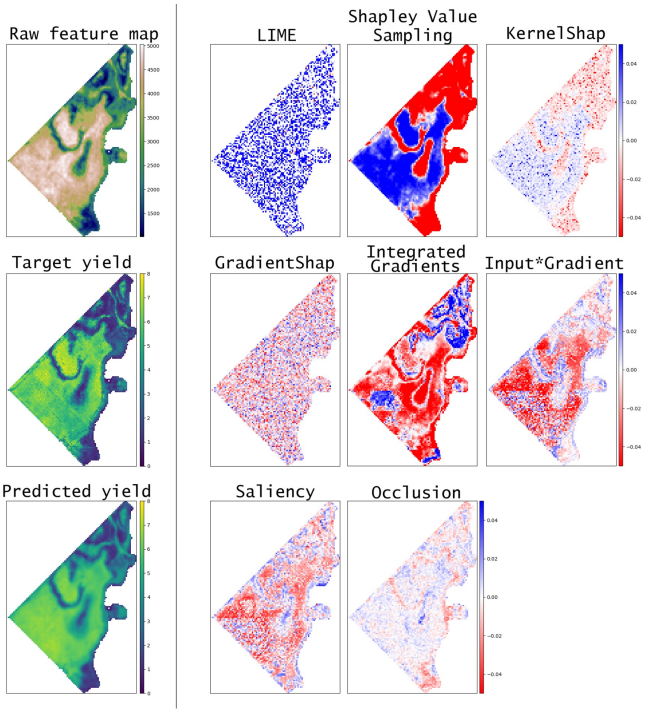


Fig. 3.5.: Qualitative evaluation of the feature attribution methods using the mean baseline on ARG-S dataset: Attribution maps of band B08 of time step 14 (March of the second calendar year).

method, it involves repetitive perturbations of the input and evaluations of the model prediction, making the interpretation of the entire datasets computationally expensive. Therefore, we select and explain the predictions of a subset of pixels, based on a random selection process in which we pre-select a maximum of five fields per farm and explain all pixels within these fields. Similar to the method comparison, we interpret the model from the cross-validation fold that achieved the highest R^2 scores across the three training scenarios (*raw-ts*, *24ts*, and *24ts+ads*).

Spectral Analysis Across the three experiments conducted for each dataset, we compare the importance of satellite bands by evaluating the spectral attribution $\mathrm{SA}(X)$ for each pixel X within the explained fields, before averaging the results. In experiments with additional modalities, we specifically rescale S2 attributions to ensure they sum up to 1, in order to facilitate the comparison against the other two experiments trained solely on the satellite time series. The results are illustrated in Figure 3.7.

Identification of the most important S2 bands

When examining which bands are frequently attributed high importance across the different experiments, we observe that each band is significant in at least one experiment, meaning that the importances are distributed over all bands. This explains why using only VIs, which do not necessarily include all available bands, results in poorer performance compared to training with the full satellite spectrum, as previously dis-

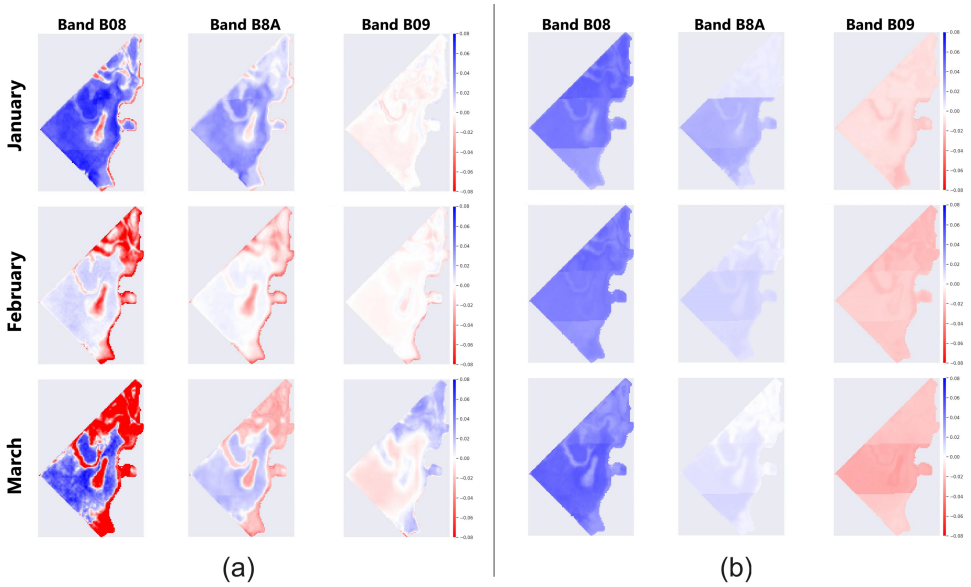


Fig. 3.6.: Comparing svs attribution maps using the mean baseline in (a) and the padded baseline in (b), over three months from the second year and three satellite bands.

Tab. 3.5.: Cosine similarity scores between the average spectral importances of different experiments.

Compared experiments	ARG-S	URG-S	GER-R	GER-W
24ts, raw-ts	0.93	0.84	0.90	0.93
24ts, 24ts+ADS	0.92	0.91	0.95	0.98

cussed in 3.2.4. Our observation further provides an argument against restricting the satellite-based data to a set of indices or RGB bands alone, a practice often carried out in other crop yield prediction studies [329, 55, 233, 201, 151].

Quantitative comparison across experiments

Comparing the raw against monthly temporal sampling strategies (i.e. *raw-ts* and *24ts*), we observe more similarities than differences between the corresponding bars across various bands and datasets. In Table 3.5, we compute in the first row the cosine similarity between these bars for each dataset. The values indicate that the band importances in Uruguay dataset are the most affected by the change in the temporal sampling strategy. To assess the impact of adding more modalities, we compare the monthly sampling experiments: *24ts* and *24ts+ADS*. We notice that in soybean datasets, i.e. in Argentina and Uruguay, the spectral importances between the two experiments are less similar than in rapeseed and wheat datasets in Germany, as reported in the second row of Table 3.5.

The case of the famous NDVI

Spectral Attributions & Vegetation Indices In the following, we connect the spectral attribution results and the model performance when trained on VIs. The Normalized Difference Vegetation Index (NDVI) is frequently used for yield prediction as highlighted in [228], and relies on the near-infrared (B08) and red (B04) bands. In the spectral importance

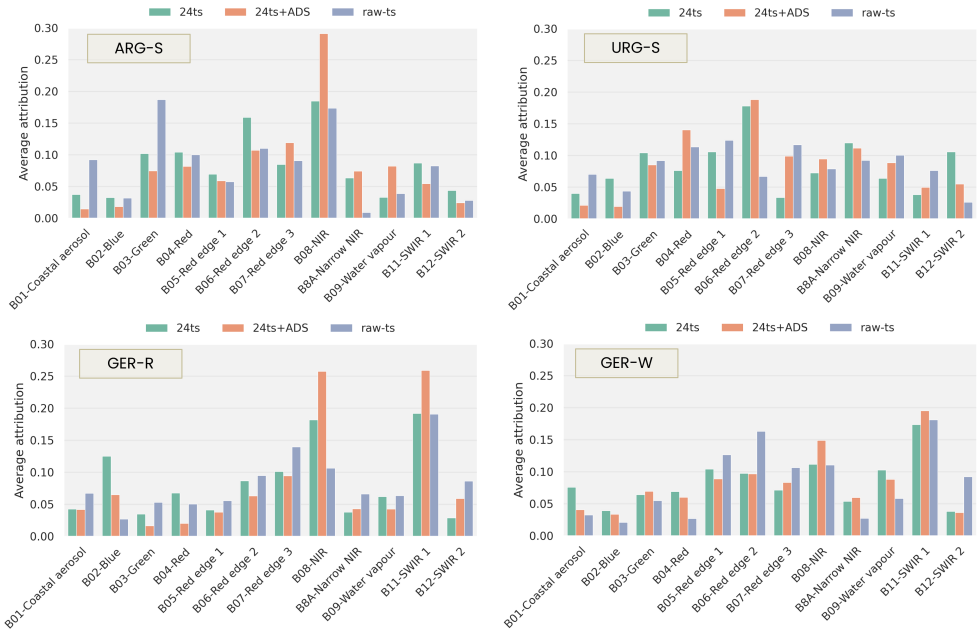


Fig. 3.7.: Average attributions per spectral band, in each dataset and under the three training settings: monthly sampling (*24ts*), monthly sampling with additional data sources (*24ts+ADS*), and raw time series (*raw-ts*).

Connecting prior results with spectral attributions

results in Figure 3.7, we notice that B08 emerges as significantly important in all datasets, aligning with its universal relevance, while B04 is assigned high attributions in soybean datasets in particular. Additionally, the S2 satellite imagery incorporates three red edge bands, of which only two are actually employed in the VIs we selected, as can be verified in Table C.2. However, the results underline that the third red edge band (B07) also exerts a high influence on the models across all experiments. Moreover, the first SWIR band (B11), also absent from the indices we used, has particularly high attributions in rapeseed and wheat datasets. This can potentially explain the substantial drop in the VI experiments for these two datasets, as reported in Table 3.3 and 3.4. The spectral importance enables the identification of the most important bands for each crop and region, and these results can efficiently be exploited to define specialized indices for the crop yield prediction task, as we will explore in Chapter 6. Combining these results with the growth stage attributions, VIs can further be exploited to estimate the length of each growth stage.

XAI-guided selection of indices

Aggregation over growth-stages

Temporal Analysis To analyze the temporal attribution results, we directly assess their aggregations over the growth stages, (i.e. $\mathrm{TA}(X)_{t_g}$ for pixel X during growth stage t_g), averaged across all the explained data points. We plot the results for the different experiments and datasets in Figure 3.8, and report their corresponding similarity scores in Table 3.6.

Results discussion

We observe that the two experiments based on monthly temporal sam-

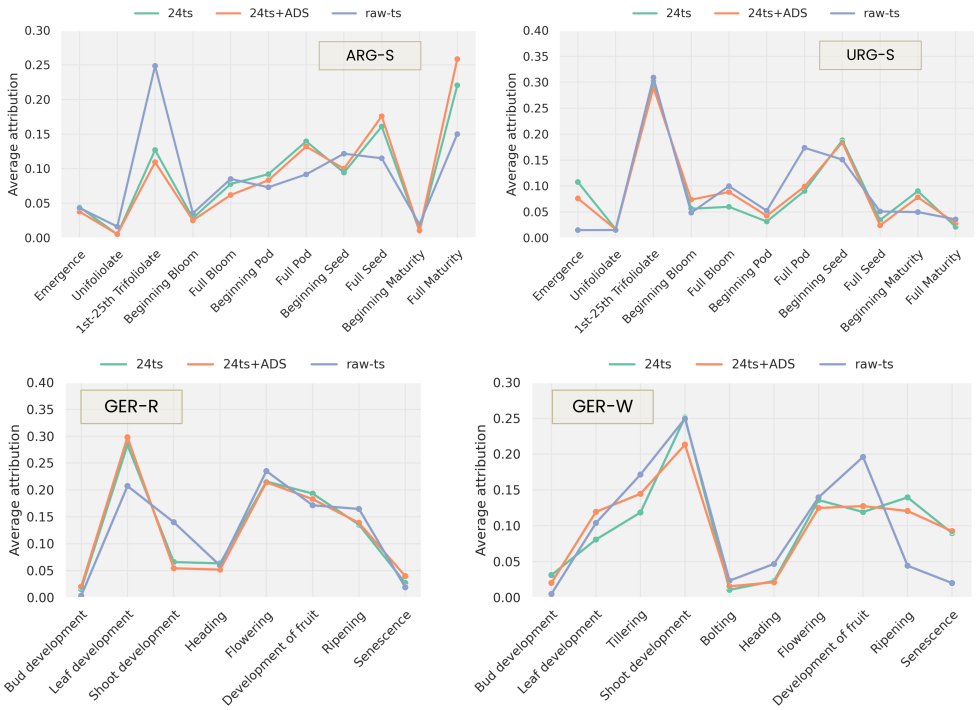


Fig. 3.8.: Average of total attributions per growth stage in each dataset, under the three training settings: monthly sampling (*24ts*), monthly sampling with additional data sources (*24ts+ADS*), and raw time series (*raw-ts*).

Tab. 3.6.: Cosine similarity scores between the average temporal attributions aggregated along growth stages of different experiments.

Compared experiments	ARG-S	URG-S	GER-R	GER-W
24ts, raw-ts	0.91	0.94	0.96	0.92
24ts, 24ts+ADS	0.99	0.99	1.00	0.99

pling, *24ts* and *24ts+ADS*, treat data from various stages similarly, since the two corresponding curves closely align across different datasets. This is further confirmed by the similarity scores higher than 0.99%, indicating that training with additional modalities have minimal impact on the significance of each growth stage to the final predictions. However, when we compare the *24ts* and *raw-ts* curves, we observe a more disparities in their values at many stages. We therefore conclude that using the raw time series of satellite data influences the importance of each growth stage in the model. This outcome aligns with our expectations, as the distribution of growth stages over time is more finely-grained than the monthly sampling, making the density of the raw time series better suited to adequately capture each growth stage. Among the stages which gain more importance when using the full S2 time series are (1) the beginning of bloom, pod and maturity in ARG-S, (2) full blooming and beginning and full pod in URG-S, (3) shoot development and heading in GER-R, and finally (4) bolting, heading and flowering in GER-W.

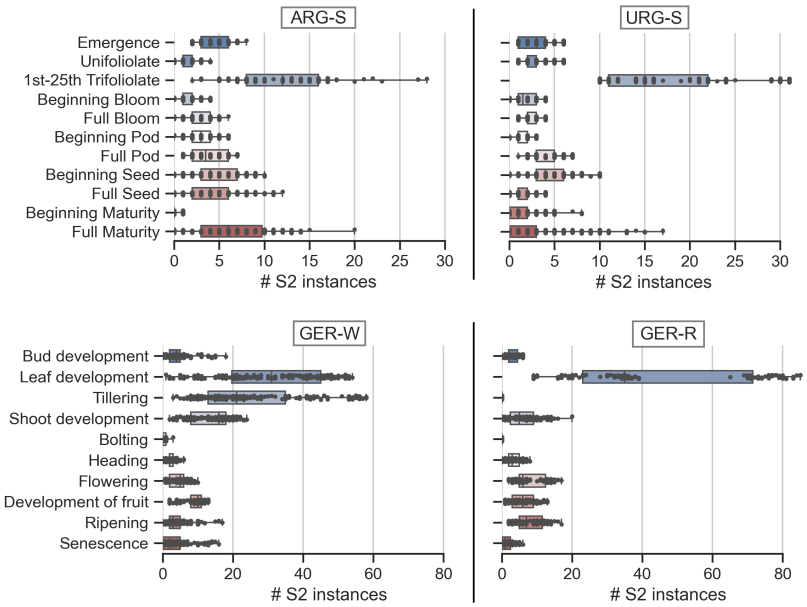


Fig. 3.9.: Number of satellite instances available per growth stage in each field of each dataset.

Influence of the growth-stage length on the aggregated attribution

To assess the impact of growth period lengths on the attribution of experiments based on the raw time series data, we count the number of satellite instances in each stage and for each field in the different datasets, and we plot the results in Figure 3.9. Upon examination of the soybean datasets, we notice that the extended duration of the third stage and the short late stages positively correlate with their respective attributions in the *raw-ts* experiment. In the rapeseed dataset, a similar pattern is observed for the early stages, while the late stages have a high importance despite their short period. In the case of the wheat dataset, the correlation is weaker, where stages such as shoot and fruit development, despite not being the longest in duration, exhibit the higher attributions. These observations lead us to conclude that while in soybean datasets, the length of growth stages and the abundance of corresponding satellite data positively correlate, such a relationship has less significance in wheat and rapeseed crops.

Conclusion on plausibility of the model reasoning

Our analysis suggests that while the duration of growth stages may impact their influence on the model, it is not the sole factor influencing their overall attribution. Hence, it is reasonable to assume that the model has captured some crop-specific characteristics rooted in agrobiology and phenology. These characteristics, in turn, shape its predictions, resulting in a close alignment of the reasoning learned by the model with actual yield factors. We further validate this assumption in the following.

Verification against domain knowledge We verify the alignment of observed patterns in the model reasoning against well-established facts in crop growth and development research. In soybean datasets, **emergence**, **unifoliolate** and **trifoliolate** stages collectively constitute the vegetation

Growth-stage significance in soybean fields

phase. The main stem nodes and their branching develop during this phase, influencing the canopy structure and the final number of nodes, as elucidated by Kumudini [170]. Since various environmental and genetic factors can impact the vegetative development, a poor canopy expansion can be identified at the last stage, the **1st-25th trifoliolate**, and linked by the model to lower yield values. The **beginning pod** stage is also strongly related to yield, with favorable temperature and moisture conditions resulting in a higher pod number per plant, more beans per pod, and larger seeds [209]. The following three stages, which extends from **full pod** to **full seed** stages, are important as the pods grow rapidly, and seed development begins, making it a crucial period for seed yield [209]. This seed-filling phase is proven to have a positive correlation with yield [101, 278], and any stress during this phase has a more significant impact on yield reduction than at other times [209]. Finally, the pronounced significance of the full maturity stage in Argentina can be attributed to the enhanced visibility of the pods. At this stage, approximately 95% of the pods on the main stem attain their mature pod color, offering a valuable visual indicator of yield, as supported by Kumudini [170]. One of the factors that could account for the differences in the results between Argentina and Uruguay is the notable influence of the flowering stage. As documented in [209, 170], soybean flowering is influenced by factors such as crop variety, day length, and temperature. Consequently, the variations observed in these regions may be attributed to differences in agricultural practices, including the application of fertilizers and herbicides.

Differences between soybean fields in Argentina and Uruguay

In rapeseed, the attribution results of juvenile growth (i.e. from **leaf development** to **heading**) is attributed to the close connection between the seed yield and the number of pods per plant, which is supported by the dry matter produced during this period. There is a linear relationship between the cumulative production of dry matter until flowering and pod density, as documented by Diepenbrock [62]. He further highlights that the **flowering stage** is the most pivotal phase impacting rapeseed yield, which aligns with the high attribution of this stage in the bottom left plot of Figure 3.8. The ability of satellite data to capture this crucial stage is attributable to the increasing flower cover, which significantly enhances photon reflectivity and absorption to 60-65% of incoming radiation. At the **fruit development stage**, pod filling is initiated and continues through the **ripening stage**. Seed yield is linearly related to photosynthetically active radiation that is intercepted during this phase, which aligns with our attribution results. Moreover, while several factors during flowering can limit the yield, rapeseed has the potential for growth after flowering, which compensates for losses of buds, flowers and pods [62].

Growth-stage significance in rapeseed fields

In the wheat dataset, significant importance is observed during the **tillering stage**, which is recognized in expert studies to have great agronomic importance in cereals, as confirmed by Acevedo et al. [2]. While it is the most important process governing canopy formation, it is also known that not all tillers produce spikes, which are the grain-bearing tip located at the top of cereal stems, as noted by Gallagher and Biscoe [94]. As **shoot development** progresses, spikes grow, the pseudo-stem

Growth-stage significance in wheat fields

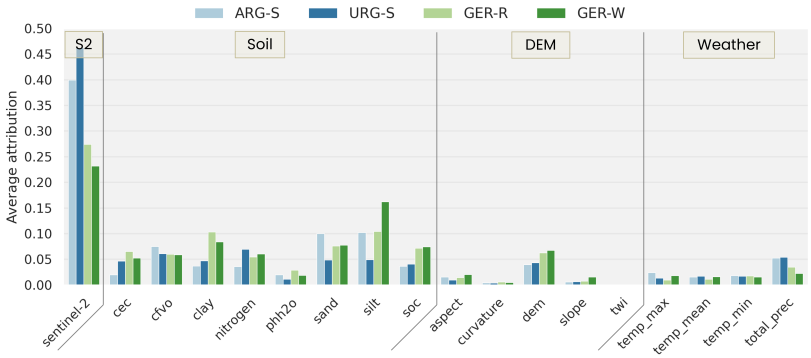


Fig. 3.10.: Average attribution of each feature in the additional modalities, contrasted with the total attribution of the satellite data across different datasets.

becomes more erect, and leaf sheaths elongate. This results in a more distinguishable canopy density at satellite resolution, aligning with the high attribution of the **shoot development stage**, as illustrated in the bottom right plot of Figure 3.8. Additionally, florets are initiated during this stage; however, it's noteworthy that less than half of these florets will complete anthesis [2]. As the crop progresses to the **heading stage**, the florets become ready for pollination and fertilization. It is at the **flowering stage** that the model leverages the blossoms for its predictions, as the number of flowers successfully pollinated directly influences the number of kernels per head. This observation accounts for the pronounced importance attributed to the **flowering stage**. During subsequent stages, the crop undergoes grain filling and moisture reduction during ripening, achieving maximum dry weight, and physiological maturity. During this period, the green color progressively fades, a change that can be captured in the satellite imagery.

Model alignment with agronomic studies

The analysis conducted above validates our temporal importance results aggregated over growth stages, strongly supporting that our models learned agronomically meaningful cues for accurate yield predictions.

Modality Attributions

Within the *24ts+ADS* experiments, we examine the total attribution of satellite data relative to each feature from the additional modalities, as depicted in Figure 3.10. We notice that while S2 accounts for nearly half of the attributions in Argentina, and around 40% in Uruguay, it contributes only about 25% of the total attribution in rapeseed and wheat datasets. The majority of the remaining attribution is distributed among the soil bands, followed by weather and terrain elevation features. In the case of the DEM data, we notice that most of the attribution is concentrated on the raw elevation values (i.e. the *dem* feature), suggesting that the additional modalities can be further filtered to use only the features relevant for the yield prediction [230].

Total attribution per feature across modalities

Total attribution per modality

We further compare the total attribution per data source in Figure 3.11, differentiating between the explained fields (i.e., the vertical bars) in each dataset. The attributions are averaged across the pixels of each field

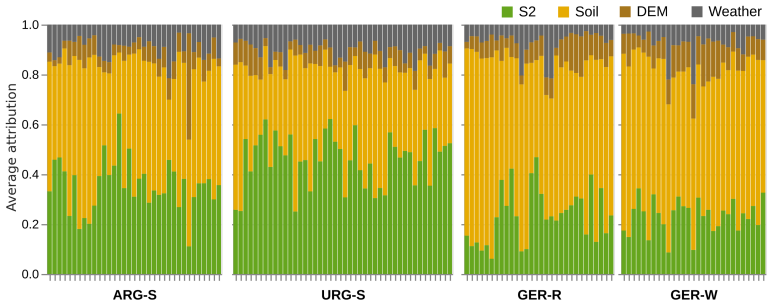


Fig. 3.11.: Total modality attribution, averaged over pixels of each explained field separately (i.e. each bar is a field), in each dataset.

separately. The area that each color covers serves as a visual indicator of the total importance each modality has in the different datasets. The yellow color, in particular, highlights the significance of soil bands in most fields and datasets, particularly in Germany, where it substantially surpasses the total attribution of S2 bands. In contrast, weather and DEM exhibit relatively lower importance across all regions, with weather being more important for soybean crops, and DEM showing more significance for rapeseed and wheat crops in Germany. It is worth noting that the literature, particularly the work of Thomson et al. [300], emphasizes the important impact of topography on wheat crop productivity, further supporting our findings.

Supplementary Analysis

In-Field variability We closely examine the attribution maps to understand how the model captures in-field variability. This is important given that our training pipeline provides the model with individual pixel data without any information about neighboring pixels, yet the model learns to approximate the variability we observe in the reference yield maps. When revisiting the attribution maps in Figure 3.6.a, we notice that some features have high variance across the field's pixels, such as band B08 in March, while others consistently show either low or high attribution values, namely band B8A in January. Hence, we exploit this characteristic to identify features (i.e. (band, time step) pairs) that effectively capture the in-field yield variability learned by the model. Specifically, for each field we calculate the variance across different pixels, then rank the features in the descending order of their variance. In Figure 3.12, we present the results for three fields in Argentina by displaying the attribution maps of the top 5 features. We observe that the patterns on the attribution maps match those in the prediction map, confirming that features with high variability are key in helping the model best capture the yield variability on the subfield level.

Connecting variability of the predicted yield with attributions at field-level

The grid of the standard deviations of the relative scaled attributions (i.e. $\breve{a}$ vectors) shown in Figure 3.13 can also be used to directly identify the features with significant influence on the model's ability to learn the in-field variability for a specific field. In this figure, we compute the attribution deviations for the field depicted at the bottom of Figure 3.12.

Identifying influential features via the variations grid

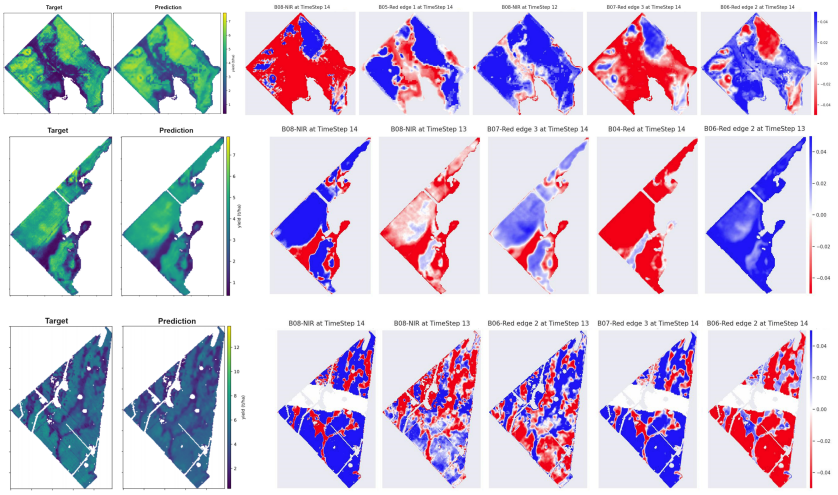


Fig. 3.12.: In-field variability of three fields from ARG-S. From left to right: target and predicted yield maps, followed by the attribution maps of five features with the highest attribution variance. (The horizontal strip on some maps from the field at the bottom are padded pixels because of cloud coverage.)

For that, we use the model trained under the monthly sampling. We observe that the influential features are primarily the NIR and Red-Edge (RE)-2 bands during the last months before harvesting. The harvesting season typically occurs around time step 15 (April of the second year) in most fields.

Attribution and performance correlation When evaluating the model on different fields, we observe important variations in the results. Therefore, we examine if a field for which the model performs poorly has a particular attribution distribution that differs from fields where the model performs better. To achieve this, we consider two types of attribution vectors: the original vector, which contains 24 time steps, or its aggregation over the growth stages. For each type, we start by defining the reference vector as the average across all considered fields. Next, we calculate the distance between this reference vector and each field's average attribution. The results of this analysis on ARG-S dataset are displayed in Figure 3.14. We observe a negative correlation with both attribution vector types, indicating that fields with a high R^2 score tend to have a closer distance to the reference distribution. To further quantify this correlation and its statistical significance, we conduct a t-test with a threshold of 5%. The results are reported in Table 3.7 (see the two rows marked with '(*all fields*)'). The p-value exceeds 6% when using the monthly attribution vectors, while it falls below 1% with the vectors aggregated over growth stages. Consequently, the latter method is more reliable in identifying fields with low performance.

Adjusting the reference vector

We further explore different approaches to define the reference vector. Instead of averaging across all explained fields, a more reliable baseline can be obtained by averaging the attributions across the best-performing

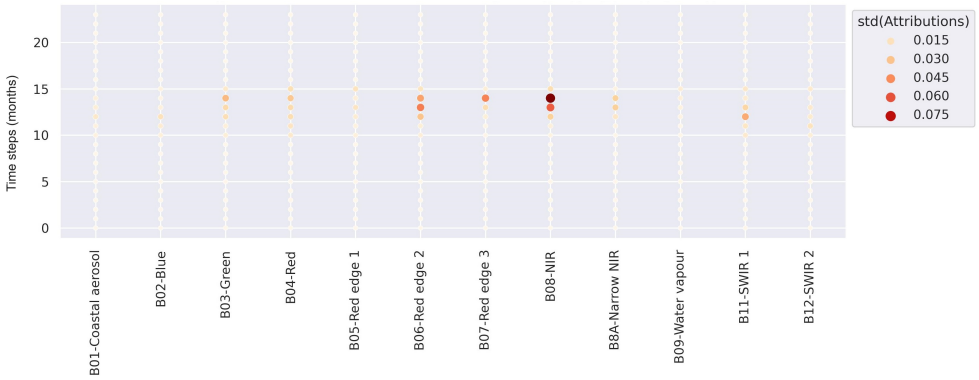


Fig. 3.13.: Standard deviation of the relative scaled attribution values across all interpreted pixels from a single field in Argentina dataset.

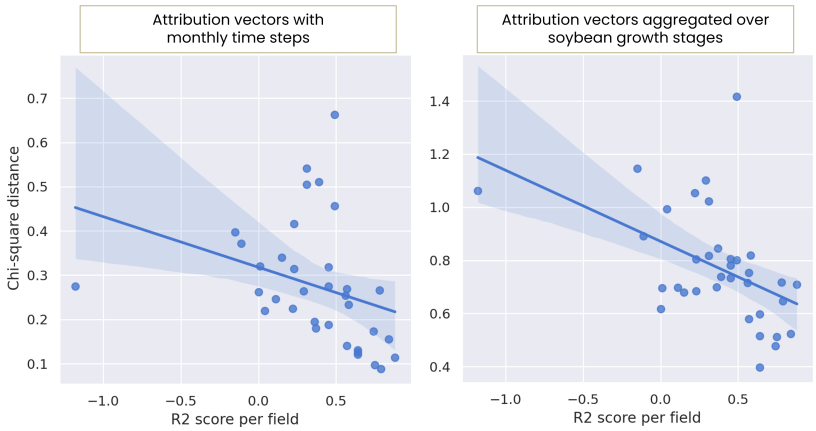


Fig. 3.14.: Comparison of the R^2 score of each field against the distance of the attribution vectors of its pixels to the reference distribution. Each point is a field in ARG-S. Here, the reference distribution is the average distribution over all explained data points in Argentina.

fields only. These are identified either by setting a threshold on their R^2 score or by specifying a predetermined number of fields with the highest R^2 scores. The corresponding experiments are included in Table 3.7. P-values below 5% are highlighted in bold. We observe that most of the tests have yielded a significant correlation, especially when using the aggregated attribution vector, which further supports its superior efficiency over the original vector in distinguishing between the fields on which the model is well- or poorly-performing.

Analysis of the model trained on raw time series

While the results described above originate from the model trained under the monthly sampling setting, we repeat the same experiments for the model trained on the raw time series, and report the corresponding results in Table 3.8. We notice in this case that comparing the raw time series of the attributions leads to low correlation scores with poor statistical significance. Conversely, when comparing growth stage aggregations of the attributions, we notice the opposite trend. We attribute this behavior to the varying sequence lengths resulting from the usage of raw time

Tab. 3.7.: Correlation and t-test results between the attribution distance of ARG-S explained fields to the different references, and their R^2 score. The attribution scores used are from the model trained under the monthly sampling. P-values below 0.05 are highlighted.

Temporal dimension	Score threshold	Best N fields	Correlation	P-value
Monthly time steps	-	(all fields)	-0.32	0.063
	-	3	-0.30	0.076
	-	5	-0.33	0.056
	-	10	-0.38	**0.025**
	-	15	-0.39	**0.021**
	0.4	-	-0.39	**0.020**
	0.5	-	-0.38	**0.024**
	0.6	-	-0.38	**0.023**
Aggregated over growth stages	-	(all fields)	-0.48	**0.003**
	-	3	-0.30	0.077
	-	5	-0.38	**0.025**
	-	10	-0.43	**0.009**
	-	15	-0.46	**0.005**
	0.4	-	-0.48	**0.004**
	0.5	-	-0.45	**0.007**
	0.6	-	-0.44	**0.008**

series data, where each time step originates from different timestamps across various fields. In contrast, comparing the attributions of growth stages proves to be more robust against these temporal disparities.

Tab. 3.8.: Correlation and t-test results between the attribution distance of ARG-S explained fields to the different references, and their R^2 score. The attribution scores used are from the model trained with the raw satellite time series.

Temporal dimension	Score threshold	Best N fields	Correlation	P-value
Raw time series	-	(all fields)	-0.17	0.324
	-	3	-0.27	0.122
	-	5	-0.22	0.205
	-	10	-0.17	0.326
	-	15	-0.17	0.336
	0.4	-	-0.19	0.267
	0.5	-	-0.18	0.291
	0.6	-	-0.18	0.296
Aggregated over growth stages	-	(all fields)	-0.37	**0.0286**
	-	3	-0.34	**0.0490**
	-	5	-0.36	**0.0358**
	-	10	-0.39	**0.0202**
	-	15	-0.39	**0.0220**
	0.4	-	-0.38	**0.0225**
	0.5	-	-0.38	**0.0232**
	0.6	-	-0.38	**0.0236**

Comparing soybean fields between Argentina & Uruguay

Soybean data analysis We further conduct a comparative analysis of the attribution scores of soybean fields from Argentina and Uruguay datasets. We anticipate overlapping patterns in the results for both datasets, as they contain the same crop in nearby regions. To this end, we make use of dimensionality reduction tools to verify the similarity between the input data and attribution results from both countries. We randomly select and combine 10,000 points from each dataset and study three vectors

from each sample: (i) the input satellite data, (ii) the attribution vector, (iii) and the attribution vector aggregated per growth stage. We then project them into a two-dimensional embedding space using Principal Component Analysis (PCA) and present the results in Figure 3.15.

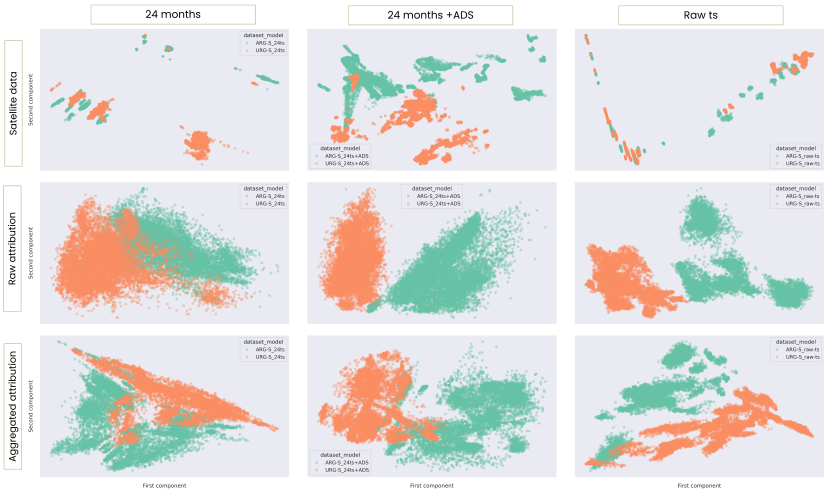


Fig. 3.15.: PCA results comparing Argentina and Uruguay data. The columns correspond to the different training settings, and the rows correspond to the type of vectors on which the dimensionality reduction is applied.

Discussion of the dimensionality reduction results

Contrary to our expectations, the plots indicate that the spectral and temporal importance show distinct patterns between the two datasets. We first examine the satellite vectors. We observe that data from the same country consistently forms multiple clusters in all of our experiments. This suggests that the source data from Argentina inherently differs from that of Uruguay. Additionally, in the second column, we can see that the two datasets are much more distinguishable when we include additional modalities in our training, as opposed to using only S2 data (see the first and last columns). This observation implies that information related to soil, weather, and terrain elevation also holds distinct patterns in the two countries.

Satellite vectors

Attribution vectors

Regarding the attribution vectors, in the second and third rows in Figure 3.15, we observe a clear separation between Argentina and Uruguay data points when training with the full raw time series. However, in the monthly sampling experiments, we observe relatively more overlap between data points from both regions. These results imply that incorporating more time steps into the satellite time series leads to the emergence of more region-specific characteristics within the data. When comparing the raw attribution vectors with their aggregation over growth stages, i.e. second and third rows, we note that the latter exhibits more overlap between the two datasets. We assume that this discrepancy is attributed to the issue of time series misalignment in the raw data, which is significantly reduced when aggregating over growth stages.

3.2.6 Summary

We summarize the explainability analysis conducted is this study in the following six major findings:

1. Evaluating multiple feature attribution methods, qualitatively and quantitatively, demonstrated that SVS is the most robust against small input variations (i.e., low sensitivity) and that the impact of these variations aligns well with their sign and magnitude (i.e., low infidelity).

2. Incorporating additional data sources improved the model performance in all datasets, had minimal effect on temporal attributions, and only a slight impact on the spectral attributions of satellite bands. Soil consistently emerged as a high-contribution modality.

3. Training with the complete satellite time series (i.e., raw sampling) outperformed the monthly sampling for all datasets at the subfield level. Additionally, the difference in temporal sampling had a greater impact on the spectral attribution scores compared to the temporal attributions.

4. Leveraging crop growth stages to interpret temporal attributions enabled the comparison between different temporal sampling strategies, facilitated the verification of model reasoning, and revealed that the stages with high contribution to the model predictions are also known to be agronomically meaningful and critical for a healthy crop growth.

5. Visualizing attribution maps for individual features at different time steps and calculating the standard deviation of each map's attributions helped identify features that enable the model to learn in-field variability of the yield.

6. Comparing the input data and attribution results for soybean fields between Argentina and Uruguay revealed significant differences in satellite and additional data, which also influenced the corresponding attribution values.

Follow-up work

Overall, the study conducted in this chapter provides valuable insights into the spectral and temporal importance of satellite data to predict crop yield. Two important directions can be further explored to extend this work. On one hand, using an intermediate-fusion based model is particularly relevant to handle the static and dynamic data sources separately before merging their representations within the network. In Section 3.3, we conduct a deep explainability analysis of such networks, and compare the performance of SVS against intrinsic interpretability methods. On the other hand, the results discussed above have identified certain features from the additional data sources which contribute minimally to the model's predictions. In Chapter 5 we leverage these findings to reduce the input features both spectrally and temporally, focusing on training the model with only the essential features while maintaining its good performance.

3.3 INTRINSIC INTERPRETABILITY WITH INTERMEDIATE MODALITY FUSION

Multimodal learning is employed in various applications that require combining modalities of distinct types and natures. However, this diversity presents challenges in aligning input modalities during pre-processing. A practical solution to this challenge is employing an intermediate fusion approach, which involves designing a neural network capable of processing each modality separately using specialized encoders. These encoders map each modality into a common representation space, facilitating their subsequent fusion. The fused representations are then passed to a prediction head for the final prediction. Given the increased complexity introduced by such architectures, this section explores various methods to enhance the interpretability of intermediate-fusion-based multimodal networks.

3.3.1 Related work

Explainability of multimodal learning

Combining and modeling multiple modalities of diverse natures often results in complex architectures and threatens the interpretability of their decision-making process [154]. Some model-agnostic feature attribution techniques, such as SHAP [200] and Integrated Gradients [294] can easily be applied to multimodal networks. Other model-specific methods leverage attention mechanisms to highlight the importance of different modalities and their interactions, yet related studies in the literature often only visualize the attention weights of certain input samples, which provides very limited insights into the more general understanding of the model [106, 306].

Attention-based explanation

Taking a closer look at the use of self-attention mechanisms to leverage the inherent model-interpretability in RS, researchers have explored this approach for several tasks, including crop mapping [162, 333, 259, 98], land cover classification [163, 210], water quality monitoring [248], and target detection [350]. However, the analysis of self-attention mechanisms to achieve model explainability is often limited in these studies, with little focus on in-depth interpretability.

Examples in RS

For instance, Khan et al. [162] compared two Transformer-based architectures for land cover classification, employing multiple post-hoc explanation techniques to elucidate the predictions. However, the self-attention mechanism was not leveraged for intrinsic model explanation. Kim et al. [163] extracted multiple attention maps from a convolutional network embedded in a satellite's on-board system, automatically identifying samples with inconsistent maps. These samples were then sent to a ground station for correction by expert annotators before being communicated back to the satellite to update the model. While this work provides a framework that uses attention maps as a tool for improving the model in a weakly supervised manner, it remains restricted to computer vision tasks and focuses primarily on local explanations (i.e., explaining individual predictions without generalizing insights across multiple samples). In another study, Xu et al. [333] trained an attention-based long short-term memory (ALSTM) network and a Transformer model for crop mapping,

and analyzed raw attention weights to provide explanations. While their analysis was primarily descriptive, further processing of the attention weights could have revealed deeper insights into the model, as we will demonstrate in our work.

Attention-based explanation for yield prediction

In the context of yield prediction, while many studies have explored the impact of the time- and region-wise drifts on the model performance [125, 126] or used attention-based models to enhance task accuracy [213, 142, 168, 249, 187, 155], we could identify only one study which has explicitly focused on explaining such models. Tian et al. [301] used an ALSTM model, which combines a LSTM network with an attention layer, to predict winter wheat yield at the county level in central China. Their input data includes an early fusion of two vegetation indices and two climate-related features. The target yield data used covers only 22 counties from Shaanxi province, with spatial and temporal resolutions of 500m and four growth stages, respectively. However, this study does not leverage the attention mechanism for inherent explainability, and relies instead on post-hoc methods.

Main research questions

In this section, we demonstrate how the attention mechanism, particularly in Transformer-based models, can be leveraged to enhance the intrinsic interpretability of multimodal networks. Specifically, we investigate the following research questions:

RQ1: Which intermediate-fusion-based multimodal network architecture performs best for the task of yield prediction, given the four modalities used in this study?

RQ2: What can the analysis of the intermediate representations reveal?

RQ3: Which method for estimating temporal attributions is most reliable?

RQ4: Can the temporal attributions provide agronomically relevant insights?

RQ5: Which modality attribution method is most reliable?

3.3.2 *Dataset*

In this section, we continue our work on the yield prediction task to serve as a testing ground for evaluating the interpretability of multimodal networks. We employ an expanded version of the dataset detailed in Subsection 3.2.2 from the preceding section. A summary of the distribution of this updated dataset can be found in Table 3.9, while Figure 3.16 visually represents its geographical distribution.

Temporal sampling

Given that the intermediate fusion approach processes each modality separately, we maintain the original temporal resolution of the satellite and weather modalities, as shown in Table 3.2. On the other hand, a spatial alignment is applied, as the data is processed on a pixel-wise basis. This implies that, for each 10x10m pixel, we need to extract information from all input modalities. Figure 3.17 illustrates the data preparation

Spatial alignment

Tab. 3.9.: Yield data description. We train different models for each country-crop pair.

Country	Crop	Years	# Farms	# Fields	# Pixels
Argentina	corn	2017-2023	21	147	1,003,133
Argentina	soybean	2017-2023	29	289	2,103,250
Argentina	wheat	2017-2022	13	61	497,651
Germany	wheat	2016-2022	6	188	306,843
Germany	rapeseed	2016-2022	6	111	306,843
Uruguay	soybean	2018-2022	10	572	2,177,206

steps for spatially aligning the modalities and target data, as well as the pixel-wise processing using the intermediate fusion model.

Data splitting

We also change the data splitting strategy, by creating training, validation and test sets, consisting of 60%, 20% and 20% of the data, respectively. Since each sample represents a pixel from a field, we grouped samples by field to ensure that the model encounters unseen fields in the validation and test splits. To maintain a consistent data distribution, we stratified the splits by year, ensuring that each split contains data from all years.

3.3.3 *Model Performance under Complex Architectures*

To address **RQ1**, we outline in the following the models used for crop yield prediction based on pixel-wise processing of spatially aligned modalities. We test various neural network architectures for intermediate fusion approach, by encoding each modality separately, before fusing the learned representations, and subsequently making the final prediction.

Modality Encoders

Encoder networks per data type

Depending on the modality's nature (static or temporal), we use different neural network architectures to encode its representation. For static modalities, such as the terrain elevations (i.e. DEM) and soil properties, we use MLPs. For temporal modalities, such as satellite and weather data, we test three different types of architectures: recurrent networks, convolutional networks, and Transformers. Each of these modality encoders is expected to produce a representation denoted as $\mathbf{z} \in \mathbb{R}^d$. Figure 3.18 depicts the different architecture types used.

a. Multi-Layer Perceptron MLPs are a type of artificial neural network where information flows from the input layer to the output layer, without any loops or cycles. MLPs extract features by learning high-level representations through layers of neurons, each performing a weighted sum followed by a non-linear activation function. In our implementation, we use two fully connected layers: the first (intermediate) layer has a dimension of $2d$, and the second (output) layer has a dimension of d, which returns the modality representation. Batch normalization and the ReLU activation function are applied after the first layer.

b. Recurrent Networks RNNs are inherently capable of handling temporal data. They process one temporal instance at a time, learning to

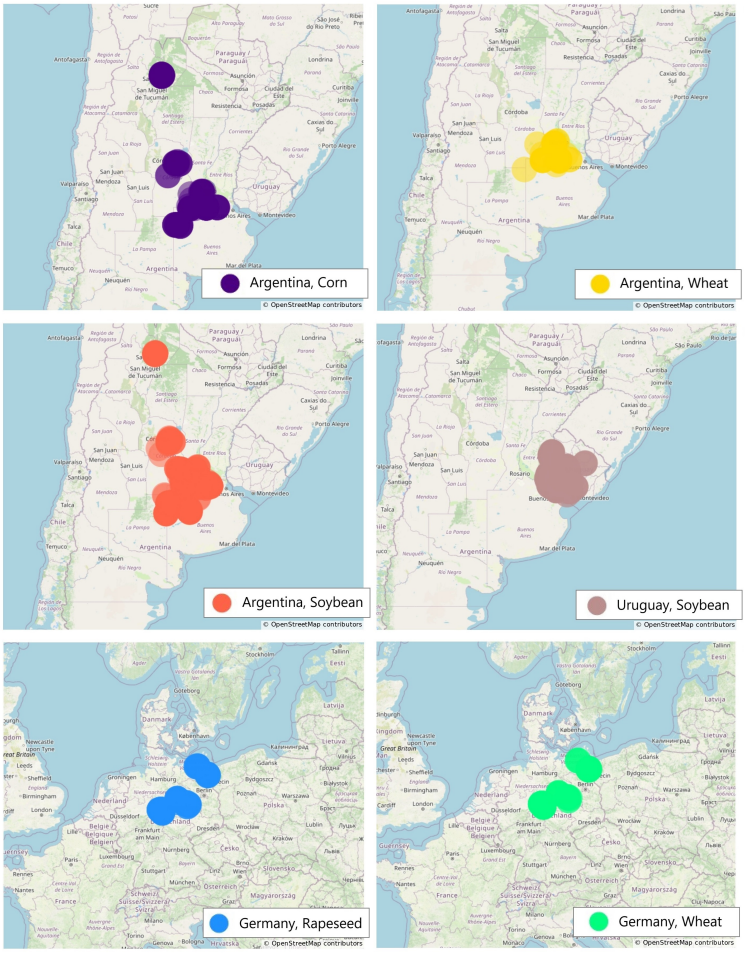


Fig. 3.16.: Geolocation of crop fields in Argentina, Uruguay, and Germany. The buffer sizes are large to protect data confidentiality. Map generated using OpenStreetMap data [236] and Plotly python package.

LSTM-based model

Attention-based LSTM

predict outputs and maintain a hidden state at each step. The hidden state is optimized to focus on important information while discarding irrelevant or redundant data. In our implementation of the RNN, we use a stack of two LSTM cells [131] with a dropout rate of 0.3, followed by a linear layer which transforms the LSTM output at the final time step to a dimension of d. Before applying the linear layer, we include batch normalization to improve training stability. We also explore another RNN , ALSTM [301], which aggregates outputs from all time steps using a weighted combination, rather than relying solely on the final time step. The weights are computed using a form of scaled dot-product attention [316].

c. Convolutional Networks 1-Dimensional convolutional neural networks (1D-CNNs) are commonly used for processing sequential data, such as time series or natural language, by applying convolutional filters across a one-dimensional input. Unlike RNNs, which process one time

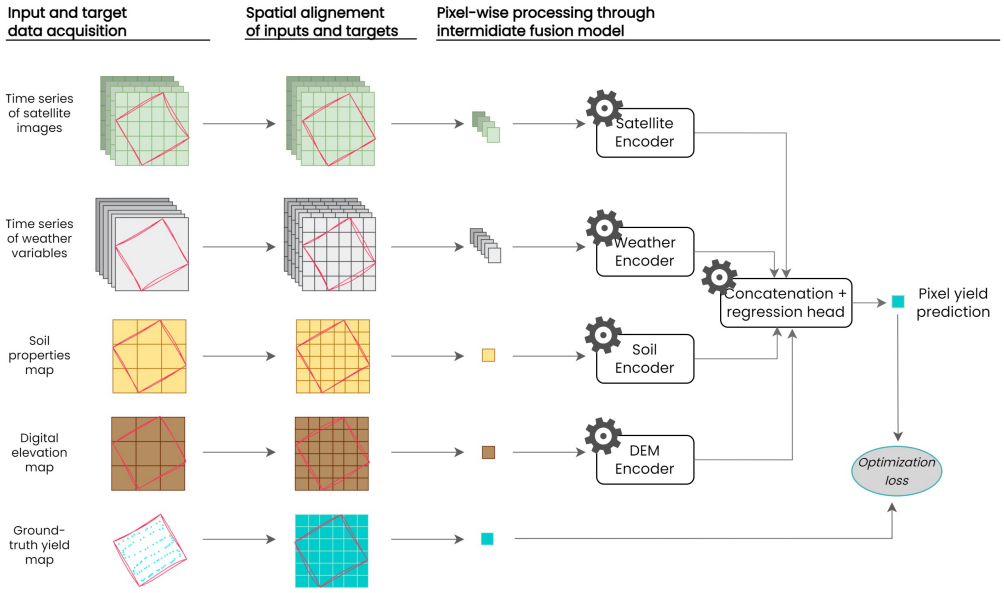


Fig. 3.17.: Diagram illustrating the overall workflow, from input modalities and target data acquisition, to model processing and pixel-wise crop yield prediction.

step at a time, 1D-CNNs use convolutional filters to capture patterns or features in a hierarchical manner along the temporal dimension of the input. Our implementation follows the feature extraction approach in [244], with the modification of using a linear layer at the end instead of a SoftMax layer to produce a modality representation of dimension d.

d. Transformers Transformers are highly effective for modeling temporal data due to their ability to use self-attention mechanisms to capture long-range dependencies within the input sequence [22, 316]. Unlike RNNs and 1D-CNNs, which process data sequentially or locally, Transformers attend to all time steps simultaneously, allowing them to more effectively capture complex temporal patterns. The input features are first passed through a linear embedding layer, which transforms each time step into a token of size d, while a learnable regression token (similar to the *class token* in [60, 69]) is added to interact with all time steps. Before adding the regression token and feeding the data to the Transformer layers, positional encoding is applied based on the date of the time step, as suggested in [316]. Specifically, we use two calendar years, covering the crop season, and for each time step, we calculate the number of days elapsed from the beginning of the first year to determine its index. This positional encoding follows the approach of [316], except we use the index calculated in days as described. The transformed input is then processed through multiple layers of Transformer encoders, each consisting of multi-head self-attention (MHA) and position-wise feed-forward networks. In each Transformer layer, the input undergoes layer normalization before being processed through the MHA. The output from the MHA layer is added back to the input via a residual connection, followed

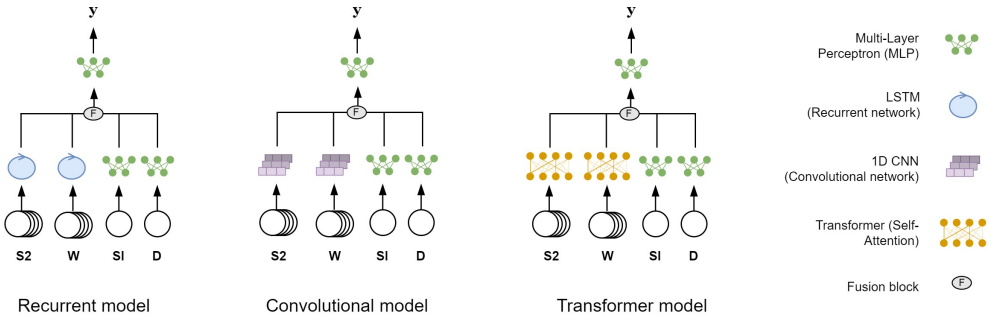


Fig. 3.18.: Multimodal architectures with concatenation fusion for yield prediction. The abbreviations used for the input modalities are as follows: S2 [Sentinel-2 Satellite], W [Weather], So [Soil], and D [DEM].

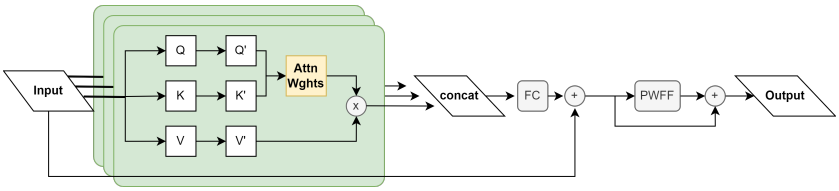


Fig. 3.19.: Schematic representation of a single Transformer layer composed of three heads. Q, K, and V stand for the query, key and values used at the attention mechanism, which are all a duplicate of the input sequence. FC refers to *fully connected layer*, while PWFF stands for the *pixel-wise feed-forward* network.

by a second normalization step. A position-wise feed-forward network is then applied, with its output also added through residual connections. This process is repeated across several Transformer layers, with the final modality representation derived from the output of the regression token. Figure 3.19 illustrates the architecture of a single Transformer layer with three heads.

e. Fusion and Regression Given the heterogeneous nature of the input modalities usually interacting in agricultural applications and RS, intermediate-level fusion is better-suited for our study, compared to early and late fusion [186, 212]. Each modality is first processed by a dedicated encoder, which maps the input into a feature representation of dimensionality d. The learned representations from all m modalities are then fused via simple concatenation along the feature dimension, and the resulting fused representation has a dimensionality of $d' = m \times d$. Finally, this fused representation is passed through a linear regression layer, which maps the d'-dimensional feature vector to a scalar output, i.e. the yield prediction.

Model Finetuning & Training

Hyperparameters selection

The different model architectures incorporate multiple hyperparameters that can influence model performance. We tested for each network

various configurations of hidden sizes, number of layers and number of attention heads (when applicable) to optimize performance for the yield prediction task. For this purpose, the dataset was split into training, validation, and test sets, with the validation set used to select the best network configuration, and the test set used to evaluate and report the model's performance on unseen data.

Training setup

The models were trained using mini-batch stochastic gradient descent with the Adam optimizer and decoupled weight decay [195]. We employed a learning rate scheduler that begins with a linear warm-up for 5 epochs, followed by cosine decay for 50 epochs [196]. Early stopping was implemented to stop training when the validation loss did not decrease for 10 consecutive epochs.

Model Evaluation

In the following, we provide answers to **RQ1**, which goal is to identify the best performing multimodal network architecture for yield prediction, and to select the model to be explain when addressing **RQ2**-**RQ5**.

a. Quantitative results To assess the performance of the models described in Subsection 3.3.3, we use the validation set to compare multiple instances, according to their R^2 score. We subsequently select the best-performing models per architecture, for which we report the test-set scores in Table 3.10, including the MAE and RMSE metrics. The subfield-level scores refer to the average performance across the pixel samples, while the field-level refers to the accuracy of predicting the average yield per field, by averaging the pixel-level target values and predictions.

Model selection

Field- and Subfield-level scores

We first observe that the subfield-level metrics are consistently worse than the field-level values. This difference arises from the increased complexity of the task of accurately predicting the yield values for individual pixels, as compared to estimating the average yield. The pixel-level predictions demand a higher level of detail and precision, making it a more challenging task than field-level predictions, where averaging helps to smooth out local variations.

Outperformance of field-level

Comparing model architectures, the Transformer model demonstrates a clear advantage in performance at the subfield-level, compared to the convolutional and recurrent networks. The scores become closer at the field-level, where the Transformer remains optimal, followed closely by the 1D-CNN. This observation aligns with the efficient performance of the temporal convolutional networks demonstrated in Chapter 5 on different EO applications. We also notice that at the field-level, the attention mechanism improves the performance of the recurrent network, highlighting the role of attention in enhancing the predictive accuracy in ML.

Outperformance of Transformers

We further compare the inference time of the best performing models in Table 3.11. Inference experiments were conducted over a batch of 1000 samples, and results were averaged across 500 batches. The machine setup included an NVIDIA V100 GPU with 16GB of memory, 50 CPUs, CUDA version 11.8, Python version 3.8.10 and PyTorch version 1.14.0. We observe in Table 3.11 that without a GPU, the convolutional model

Inference time

CPU performance

Tab. 3.10.: Comparison of model performance evaluated on the subfield-level (i.e. pixel level) and the field-level on the test set. R^2 has no unit, while RMSE and MAE are in t/h.

Model	# Parameters	Subfield-Level			Field-Level		
		R^2	RMSE	MAE	R^2	RMSE	MAE
1D-CNN	6,333,377	0.42	2.58	1.94	0.74	**1.41**	**1.02**
LSTM	54,977	0.41	2.61	2.01	0.71	1.71	1.32
ALSTM	38,017	0.41	2.59	1.99	0.74	1.47	1.19
Transformer	147,073	**0.52**	**2.35**	**1.74**	**0.78**	1.42	**1.02**

Tab. 3.11.: Comparison of model inference time on a batch of 1000 samples.

Model	CPU		GPU	
	Mean(s)	STD (s)	Mean (s)	STD (s)
1D-CNN	0.549	0.090	0.006	0.013
LSTM	4.048	0.575	0.003	0.002
ALSTM	2.800	0.482	0.003	0.004
Transformer	2.055	0.156	0.020	0.002

exhibits the highest inference speeds, slightly exceeding half a second per batch, on average, followed by the Transformer model which slightly exceeds two seconds. In contrast, recurrent networks demonstrate the slowest speeds, primarily due to their sequential processing of time steps. When using a GPU, the speed ranking is reversed, as the recurrent networks achieve the fastest processing times. Nevertheless, the inference times of all models remain mostly below 22 milliseconds.

GPU performance

b. Transformer configuration To investigate the behavior of different configurations of the Transformer-based model, we report the evaluation metrics on the validation and test sets across various model setups, focusing on the five best-performing instances, as shown in Table 3.12. We observe that the models achieve comparable performance on the validation set: the R^2 scores range between 0.75 and 0.77, the MAE falls within 1.35 and 1.43 t/ha, while the RMSE values are within the range of 1.85-1.92 t/ha. However, a relatively larger variance is observed in the test set results. The range of R^2 values varies between 0.39 and 0.52, while MAE ranges between 1.74 and 1.97 t/ha, and RMSE ranges from 2.35 to 2.64 t/ha. The models ranked first and fifth on the validation set achieve the best performance on the test test, with a moderate margin above other models. These observations suggest that the generalization capacity of the model on the validation set does not necessarily transfer to the test set. The similar performance observed in the validation set further indicates that changes in model parameters, such as the number of heads or layers, have a relatively minor impact on overall performance.

Validation set performance

Test set performance

Based on these results, and in order to prioritize simplicity and ease of interpretation over marginal gains in evaluation metrics, the explanation experiments in the following subsections will focus on the fifth model,

Model selection for explainability

Tab. 3.12.: Comparison of Transformer models performance on the subfield-level. Best and second-best scores are highlighted in bold and underlined, respectively. R^2 has no unit, while RMSE and MAE are in t/h.

Hidden size	Heads	Layers	Validation set			Test set		
			R^2	RMSE	MAE	R^2	RMSE	MAE
64	4	2	**0.77**	**1.85**	**1.35**	**0.52**	**2.35**	**1.74**
32	2	2	0.76	1.86	1.36	0.39	2.64	1.97
64	1	2	0.75	1.91	1.38	0.40	2.61	1.95
32	2	6	0.75	1.91	1.42	0.44	2.54	1.93
32	1	4	0.75	1.92	1.43	0.48	2.44	1.83

which we will refer to as the "*selected*" model. Achieving the closest performance to the best model on the test set, the selected model offers the advantage of using a single head in the Transformer encoders. The decision to select one attention head has several advantages. It simplifies the estimation and interpretation of attention-based attribution results, which we will present in detail later. By focusing on a single head, we avoid having to interpret multiple heads individually or average their results, which could lead to oversimplification or suppression of unique patterns learned by each head. Averaging results may mask important insights gained through specific heads, as discussed by Abnar and Zuidema [1] and Chefer et al. [42]. However, interpreting individual heads adds complexity to model interpretability, as mentioned by Voita et al. [320].

Model selection benefits

c. Qualitative results To visually compare the performance of the selected model from each architecture type, we selected two corn fields from the validation set, referred to as Field-A and Field-B, and plot in Figures 3.20 and 3.21 their corresponding target, prediction, and relative error maps. These visualizations help illustrate how well the models capture spatial patterns in yield predictions and highlight areas where they may struggle. The fields were chosen based on the evaluation of the selected Transformer model: Field-A, where the model performed well, and Field-B, where the model performed poorly.

Visual comparison of architectures

Visualized fields

The results for Field-A are shown in Figure 3.20. The top row displays the target yield values alongside the predicted values from the best-performing 1D-CNN first, followed by the LSTM, the ALSTM, and finally the selected Transformer model. The average yield is indicated above each map, alongside the Bhattacharyya metric scores, which measures the captured yield variance. The second row shows the corresponding error maps for each model. We notice that the Transformer model provides the most accurate approximation of the average yield, with 9.18 t/ha compared to the ground truth of 9.25 t/ha, in addition to capturing the highest yield variance, as indicated by its high Bhattacharyya score of 89.8%. The LSTM model follows closely, achieving the same average relative error of 4% as the Transformer model. In contrast, both the 1D-CNN and ALSTM model struggle to capture the in-field yield variance, as they

Results of Field-A

Comparison across architectures

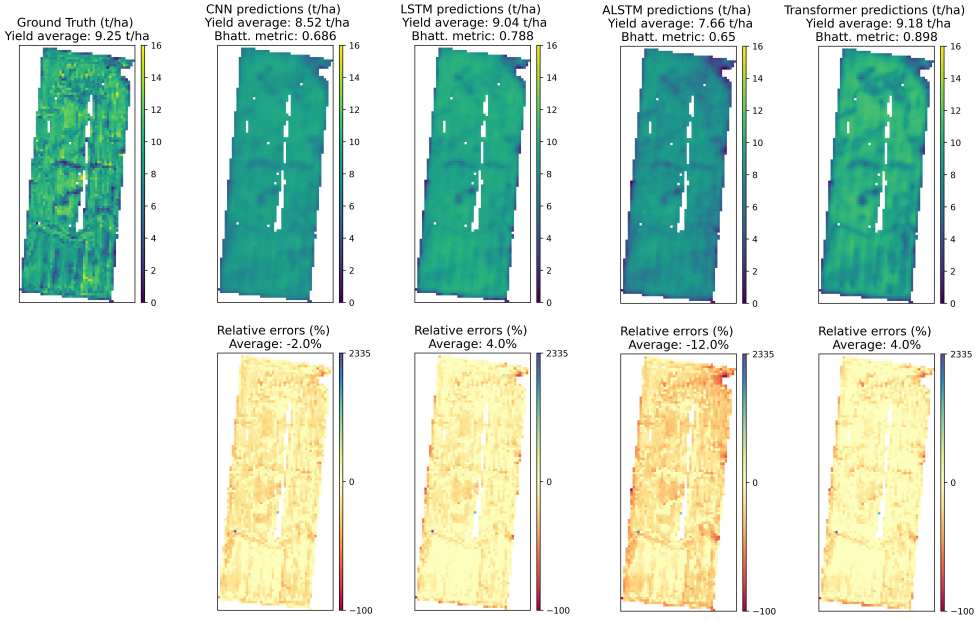


Fig. 3.20.: Qualitative results on Field-A. From left to right: ground-truth and predicted yield values of the 1D-CNN, LSTM, ALSTM, and Transformer models, with the relative prediction error displayed at the bottom.

achieve lower Bhattacharyya scores of 68.6% and 65%, respectively. The ALSTM model particularly underestimates the yield at several pixels and regions of the field, as shown in the corresponding relative error map.

Results of Field-B

Figure 3.21 presents the results for Field-B, where the Transformer model did not perform very well. The 1D-CNN and Transformer models demonstrate a relatively higher ability to approximate the average yield and capture its variance, with the convolutional model slightly outperforming the Transformer model. In contrast, the LSTM and ALSTM models achieve lower accuracy in predicting the yield, on average, and tend to underestimate the yield across many pixels, as depicted in the respective error maps. Interestingly, most models face challenges in accurately predicting the yield near the field borders, with the ALSTM model being an exception, showing better performance in these regions.

Overall, Transformers outperform

Overall, despite the challenges in Field-B, the Transformer model maintains comparatively good performance relative to the other models. Considering CPU and GPU inference times, performance improvements, and interpretability, we believe the Transformer model offers a well-balanced choice for the subsequent interpretability analysis.

3.3.4 *Interpretability of Learned Representations*

In this section, we address **RQ2** and evaluate the information content of intermediate model representations using linear probing, focusing on the selected Transformer-based model. Next, we analyze the attention weight matrices learned by the model, evaluating their similarity across pixels

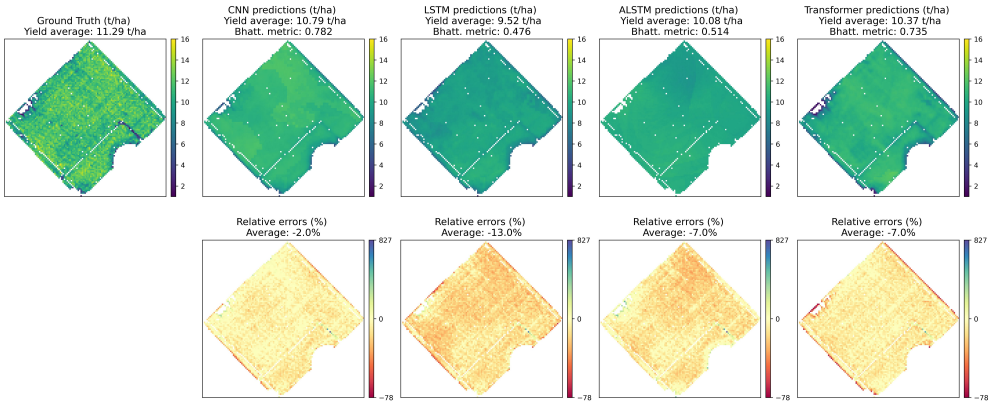


Fig. 3.21.: Qualitative results on Field-B. From left to right: ground-truth and predicted yield values of the 1D-CNN, LSTM, ALSTM, and Transformer models, with the relative prediction error displayed at the bottom.

within the same field, and examining how these weights are distributed across the different layers of the Transformer encoders.

Linear Probing

To better understand the roles and dynamics of the intermediate layers, we use linear classifier probes [8]. In practice, linear probes consist of linear regressors that take as input the latent features learned by an intermediate layer of the trained model and learns to predict the corresponding yield value, as predicted by the model. High accuracy of this regressor suggests a linear separability of the features at the examined layer. By comparing the accuracy of linear probes across successive layers, we can verify how the learned features gradually become more separable across different modality encoders.

Experimental setup

We investigate the linear separability of the intermediate layers of the selected Transformer model, i.e. the fifth model in Table 3.12. To facilitate this analysis, we randomly select 100,000 samples, representing approximately 10% of the corn dataset, using 90% of these samples to train linear probes and the remaining 10% for testing. For each layer, we compute its output given the selected samples as inputs, flatten these latent representations, and then use them to train a linear model to predict the model's final yield prediction. The RMSE scores on the test set are presented in Figure 3.22.

Results & discussion

We observe that the intermediate representations learned for the S2 satellite data demonstrate the highest linear correlation to the predicted yield values across all layers, followed by the weather data. In contrast, soil and DEM data show significantly lower linear separability. Given the static nature of these two modalities, they were processed using shallow MLPs, and they also have low spatial resolution, which contributes to their limited potential to predict the yield. When comparing the temporal modalities, i.e. satellite and weather data, the results indicate that the linear separability of weather data remains nearly constant throughout

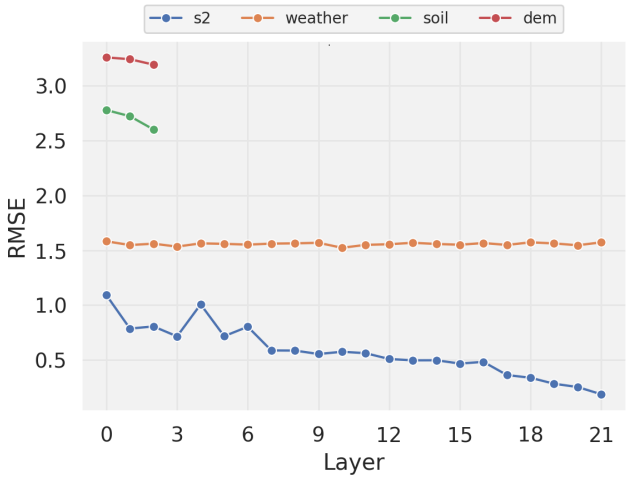


Fig. 3.22.: RMSE (t/ha) scores of the linear probes attached to the modality encoders.

the Transformer layers, whereas a significant increase is observed across the satellite encoder layers. This trend can be attributed to the higher complexity of the satellite time series, which has the highest spatial resolution and comprises 12 spectral bands, in contrast to the four weather properties used.

Attention Weights and Aggregations

We introduce in this subsection three different techniques to leverage the attention mechanism to explain the model inner reasoning.

a. Raw attention weights Since the introduction of attention mechanisms in the literature, many have seen the opportunity to use the weights for explaining neural networks in EO applications [316, 259, 333]. Indeed, the attention weights link the input to the subsequent layers of the network, allowing the model to focus on relevant parts of the input, and this link is used to interpret the model reasoning behind individual predictions.

b. Temporal attentions Temporal attentions are extracted from the attention weights to identify the time steps prioritized by the model as it makes its final prediction. Let $A^l \in \mathbb{R}^{T\times T}$ denote the attention matrix of layer l, where T is the number of time steps and $A^l_{j,t}$ is the attention weight assigned to time step t by time step j. To get temporal attentions $S^l = \{S^l_1, S^l_2, \cdots, S^l_T\}$, we first compute the total attention received by each time step t at layer l from all time steps. This value is then normalized by dividing by the total number of time steps T, resulting in a probability distribution:

$$S^l_t = \frac{1}{T}\sum_{j=1}^{T} A^l_{j,t}, \quad \forall t \in \{1,\dots,T\}. \tag{3.3}$$

This is not to be confused with the summation over each row of the attention matrix, where the weights form a probability distribution over time steps due to the row-wise SoftMax normalization [316], and thus:

$$\sum_{t=1}^{T} A_{j,t}^{l} = 1, \quad \forall j \in \{1, \ldots, T\}.$$

The time series S^l of temporal attentions is computed for each layer and head of the Transformer encoders to understand how attention is distributed throughout the network. For the final layer L, only the attention weights associated with the regression token r are evaluated, as all other time steps are excluded from subsequent processing within the model:

$$S_t^L = A_{r,t}^L, \quad \forall t \in \{1, \ldots, T\}.$$

Shannon entropy to measure information content

To evaluate the information content within temporal attentions, we use Shannon entropy as defined in the foundational work by Shannon [269]. The entropy is computed for each time series of temporal attention S^l at every layer of each Transfromer encoder. Given that the temporal attention values are normalized in Equation 3.3, we consider their range within [0,1] and divide this interval into 100 bins for computing the entropy. This ensures comparability across modalities. Low entropies indicate that the model focuses its attention on specific time steps, while high entropies suggest it spreads attention more evenly along the temporal dimension.

c. Attention Rollout (AR) In a multi-head multi-layer Transformer block, each sample generates multiple attention weight matrices. Direct analysis of each matrix can be time-consuming and might not easily reveal the inner workings of the model. Additionally, as we progress through deeper layers of the model, the identifiability of individual time steps decreases, resulting in increasingly mixed information. Consequently, direct probing of attention weight matrices for explainability becomes impractical. Therefore, to trace the information propagated from the input layer to the final embeddings of each Transformer block, we employ Attention Rollout (AR) [1]. This method treats attention weights as proportion factors and iteratively multiplies the attention weight matrices of the multiple attention layers. The resulting matrix encodes the attention distributions of the entire Transformer block and can thus serve as a reliable basis for explanation.

Limitation of raw attention weights analysis

Aggregation through matrix multiplication

d. Generic Attention (GA) Another approach that leverages the internal workings of the Transformer model and facilitates its interpretation is Generic Attention (GA) [42]. Unlike AR, which only uses the attention matrices, GA propagates information backward from the final output through the last Transformer layer and subsequently through all preceding layers, using gradients.

Gradient-based aggregation

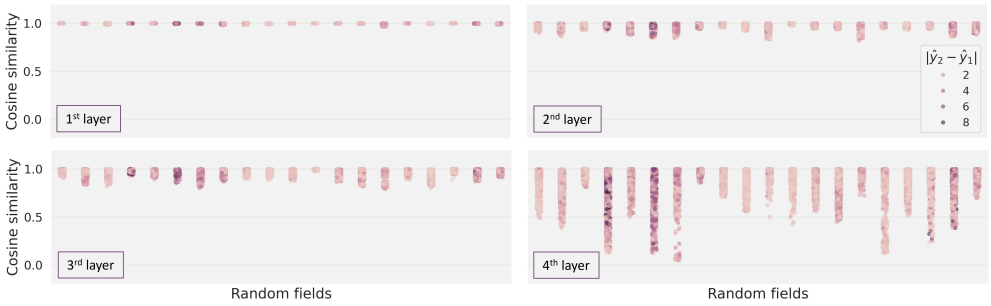


Fig. 3.23.: Cosine similarity of the attention weights from the satellite encoder of multiple pairs of pixels in a consistent set of 20 random corn fields, and the corresponding difference in prediction.

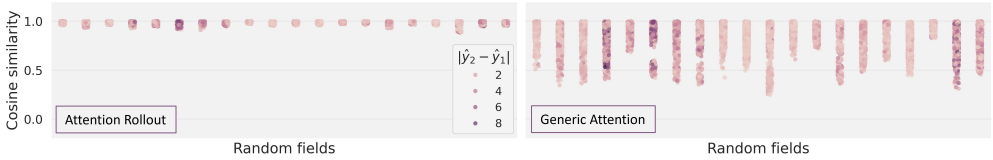


Fig. 3.24.: Cosine similarity of the AR and GA of the satellite encoder of multiple pairs of pixels in a consistent set of 20 random corn fields, and the corresponding difference in prediction.

Attention Weights: In-field Distribution

In this subsection and the following, we investigate the attention weights learned by the selected Transformer model at different levels of the satellite and weather encoders.

Weights similarity at field level

Considering that for pixels within the same field yield variations are expected to be minimal and growth conditions are similar, we quantify the similarity of attention weights at the field-level to later aggregate the attention-based explanations at this level. We also examine the correlation of attention weights similarity with the prediction similarity of each pair of pixels. This analysis is conducted through the following steps: First, 200 pixels are randomly selected from each field. Then, for each pair of pixels we calculate (i) the cosine similarity between attention weights and (ii) the difference in predicted yield, separately in each field. Finally, scatter plots are generated, where the similarity values are plotted per field and colored according to the corresponding difference in predicted yield.

Weights similarity results

An example in Figure 3.23 illustrates the results from each layer of the satellite Transformer encoder from 20 random corn fields. For the first three layers, the distance between the flattened full attention weight matrices is compared, whereas for the final layer, only the weights attending to the regression token are considered. We notice a pronounced similarity in the three first layers, but it diminishes significantly in the fourth layer in most fields. Additionally, no correlation is visually identified between the absolute prediction error and the distance between the attention weights of the compared pixel pairs. We verify quantitatively the weak correlation using the Spearman correlation metric, as reported in

Tab. 3.13.: Minimum, mean, and maximum Spearman correlation values between pairwise cosine similarity scores and prediction differences across 20 corn fields.

	Attention Weights Layers				AR	GA
	1st	2nd	3rd	4th		
Min	-0.13	-0.17	-0.10	-0.18	-0.14	-0.14
Mean	0.00	-0.01	0.00	0.00	-0.01	0.00
Max	0.09	0.09	0.09	0.13	0.13	0.12

Table 3.13. The results imply that similar predictions are not necessarily associated with a similar distribution of attention across different time steps, even for pixels within the same field.

Similarity results for AR and GA

We also conducted the same analysis to compare the AR and GA results. As shown in Figure 3.24, a strong similarity is noted between the AR attributions at the field-level, in contrast to larger differences observed in the GA results. We believe that the high similarities observed in the first three layers in Figure 3.23 should not be entirely outweighed by the decreasing similarities in the last layers, which suggests a higher reliability of AR compared to GA. Additionally, a desirable property of attribution methods is low sensitivity, meaning that minor variations in input feature values should not lead to significant changes in the attributions [336], as we have seen in Section 3.2.1. Since pixels from the same field typically experience similar environmental conditions, their input values are expected to be comparable, and consequently, their attributions should exhibit consistency as well. Furthermore, the inclusion of gradients in the computation of GA could contribute to its high sensitivity, as shown by Ghorbani et al. [105] in other gradient-based attribution methods.

Implications on sensitivity property

For the weather Transformer encoder, we observe perfect similarity across all evaluated fields, irrespective of the method used. This is attributed to the low spatial resolution of weather data, leading to identical input weather values for all pixels within the same field.

Attention Weights: Layer-wise Distribution

After assessing the similarity of the attention weights across different pixels, we now study their temporal distribution across different layers. We compute the time series S^l of temporal attentions for each layer l as described in Section 3.3.4.b. Figure 3.25 presents these results for the temporal modalities, with Field-A shown in the top row and Field-B in the bottom row. Results from other fields are displayed in Appendix E.1.

Satellite temporal attentions

In the case of the satellite time series, we observe that the attention weights from the first layer are distributed smoothly across the entire time series. In contrast, the second and third layers exhibit more peaks, which become significantly more pronounced in the fourth layer. These differences across layers were also observed in similar previous studies [333]. Moreover, the varying patterns of attention distribution across different fields confirm that each layer is capturing unique temporal dynamics relevant to the conditions of each field. For the weather encoder,

Weather temporal attentions

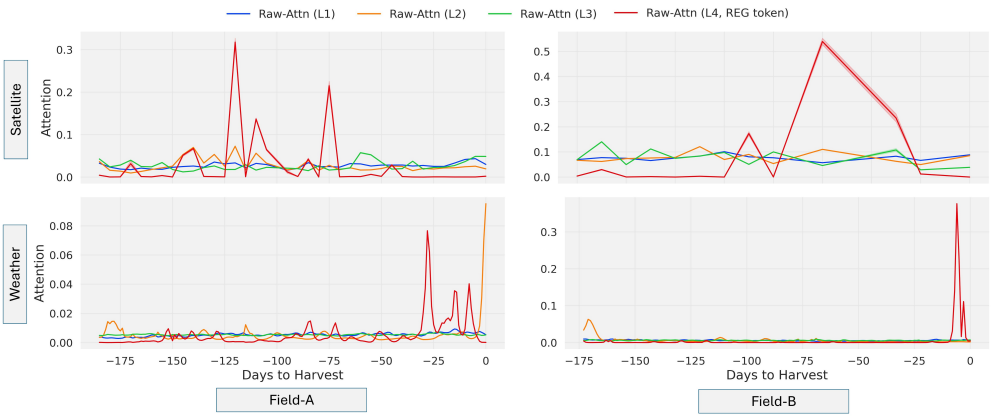


Fig. 3.25.: Total attention weights attending at each time step for the first 3 attention layers, and the regression token weights in the final layer. The results are averaged across 200 randomly selected pixels from Field-A, at the top, and Field-B, at the bottom, and are displayed for the satellite (a) and weather (b) Transformer encoders. The light buffer regions represent the 95% confidence interval around the average value.

the attention distribution results reveal that the second and fourth layers exhibit a particularly discriminative behavior across different time steps.

Shannon entropy in Fields A and B

To understand the information content within temporal attentions, we use Shannon entropy as described in Section 3.3.4.b. The results are presented in Table 3.14 for Field-A, Field-B, and three randomly selected fields (the same fields used in Appendix E.1). We observe that the first and last layers of the satellite encoder have the lowest entropy values, which indicates that most of the attention mass is concentrated at few time steps. For the weather encoder, the lowest scores are observed at the first and third layers.

Entropy across corn dataset

To gain a broader understanding of the entropy distribution across the corn dataset, Figure 3.26 visualizes entropy scores across all corn fields. The figure confirms the general observation that the lowest entropy scores occur in the first and last layers of the satellite encoder, and in the first and third layers of the weather encoder. Additionally, a comparison between the two encoders reveals that the entropy values in the weather encoder layers are generally lower than those in the satellite encoder layers. This suggests that weather information relevant for model prediction is concentrated within fewer time steps, whereas the useful satellite information is distributed more evenly throughout the growth period.

These findings highlight the differential use of attention mechanisms across modalities and how different layers of the Transformer model specialize in capturing various temporal aspects of the data, providing insights into how the model interprets and prioritizes different parts of the time series for yield prediction.

Tab. 3.14.: Entropy of temporal attentions retrieved from the satellite and weather encoders, averaged across 200 pixels from each field. Lowest entropies per field and modality are highlighted.

		Field-A	Field-B	R.Field-1	R.Field-2	R.Field-3
Satellite encoder	1st layer	1.58	2.16	**0.88**	**1.48**	**1.34**
	2nd layer	2.11	2.50	1.97	1.72	1.86
	3rd layer	2.01	3.04	1.65	2.03	2.09
	4th layer	**1.54**	**1.77**	1.50	1.50	1.68
Weather encoder	1st layer	**0.00**	**0.00**	0.39	**0.00**	**0.00**
	2nd layer	0.38	0.57	0.77	0.53	0.65
	3rd layer	**0.00**	**0.00**	**0.25**	**0.00**	**0.00**
	4th layer	0.77	0.34	0.86	0.72	0.63

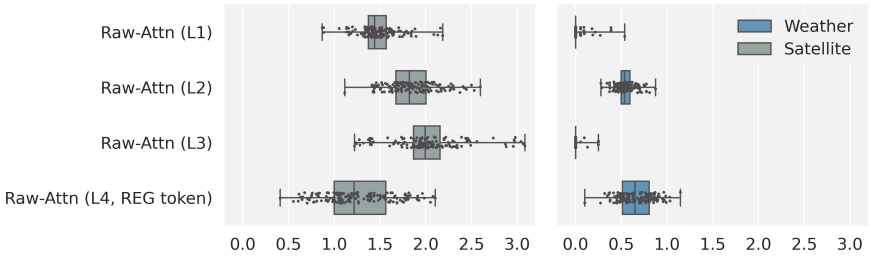


Fig. 3.26.: Shannon entropy of the satellite and weather temporal attentions, averaged across 200 pixels from each corn field.

3.3.5 *Interpretability through Input Attributions*

Methodology

Attribution methods

To address **RQ3**, we extract temporal attributions from the AR and GA attention matrices by using the attention weights of the regression token from the last layer of the Transformer block. We further include the SVS method in the analysis, to compare the two attention-based attributions against post-hoc, model-agnostic attributions. A quantitative evaluation of the three methods is conducted, using the sensitivity and infidelity scores. The SVS method and the attribution evaluation metrics are described in the previous chapter, in Section 3.2.1.

Attribution evaluation methods

Time-wise occlusion in SVS

The computation of the SVS scores involves occluding features to assess the impact of their absence on the prediction. In this chapter, we mask entire time steps instead of individual features, replacing them with baseline values computed as the mean of each masked variable across the dataset.

Temporal Attributions

We analyze in this section the temporal attributions provided by the three attribution methods: AR, GA, and SVS. Figure 3.27 displays the average attributions for Field-A and Field-B, while results for additional corn fields are provided in E.2.

a. Entropy Analysis We conduct again an entropy analysis to quantify the information compressed within the temporal attributions of each

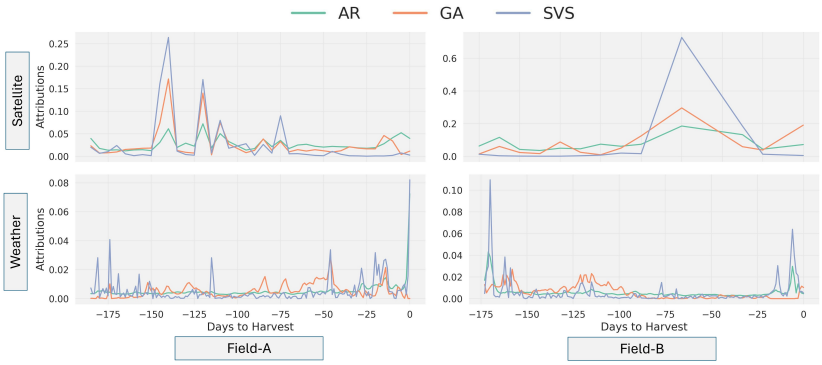


Fig. 3.27.: Field-level average attributions of the satellite and weather modalities, for Fields A and B. Due to the high computational cost associated with the SVS method, we limited the number of pixels sampled per field to 32 pixels.

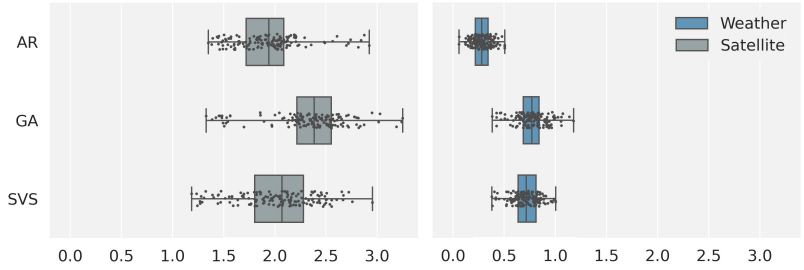


Fig. 3.28.: Shannon entropy of the satellite and weather temporal attributions, averaged across 32 pixels from each corn field.

Comparing modalities

modality. The results, shown in Figure 3.28, reveal that the attributions of the satellite data exhibit higher entropy compared to those of the weather data. This indicates that the important information in the satellite modality is distributed along a wider range of instances. Interestingly, this observation aligns with the findings from the entropy analysis of the attention weights, suggesting that the entropy patterns of the attention mechanism are preserved in the estimated feature attributions. Comparing the three methods, AR demonstrates the lowest entropy scores on average across both modalities, whereas GA exhibits the highest scores. This indicates that AR identifies a smaller subset of instances with significant influence on the model predictions. In contrast, GA identifies a broader set of important time steps. The behavior of SVS falls between these two patterns.

Comparing attribution methods

Qualitative analysis

b. Similarity Analysis A visual assessment of the similarity between the curves in Figure 3.27 further reveals a degree of alignment among the attribution methods. For instance, there is a clear correspondence among the peaks identified by the three methods within the satellite attributions of Field-A, while a (less pronounced) alignment can be noticed in the weather attributions of Field-B, particularly between AR and SVS. To quantitatively assess the similarity between the different methods, we

Quantitative analysis

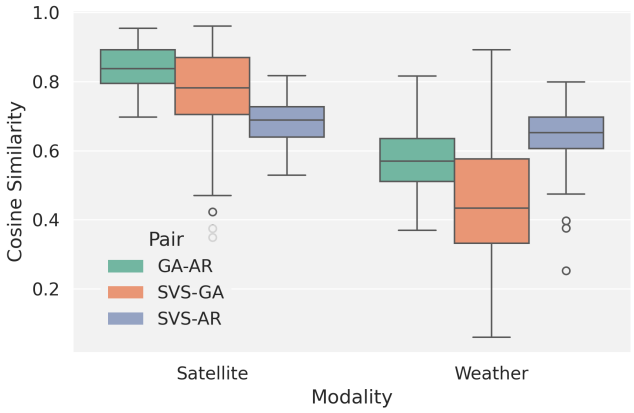


Fig. 3.29.: Distribution of field-level cosine similarities between every pair of the compared attribution methods: GA, AR and SVS.

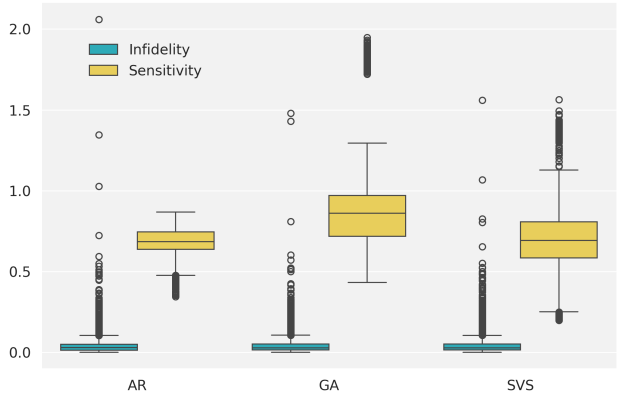


Fig. 3.30.: Infidelity and Sensitivity scores of the temporal attributions estimated by AR, GA, and SVS methods.

calculate the cosine similarity between each pair using the field-level averaged attributions, and display in Figure 3.29 the distribution of the results across all corn fields, using 32 pixels per field. When comparing the modalities, we observe that the attribution scores of the satellite data consistently exhibits higher similarity scores than weather data. This implies that the attribution methods align more closely when estimating temporal attributions for the satellite signal. A potential explanation lies in the lower entropy scores of weather data, which limits the temporal spread of important information and subsequently decreases the likelihood of alignment between methods. When comparing the methods, GA exhibits higher similarity to AR than SVS, on average. Interestingly, AR and SVS provide the most similar weather attributions and the least similar satellite attributions. Yet, the similarity scores for this pair remain within a comparable range across both modalities, indicating a more robust and consistent alignment between AR and SVS.

Comparing across modalities

Comparing across methods

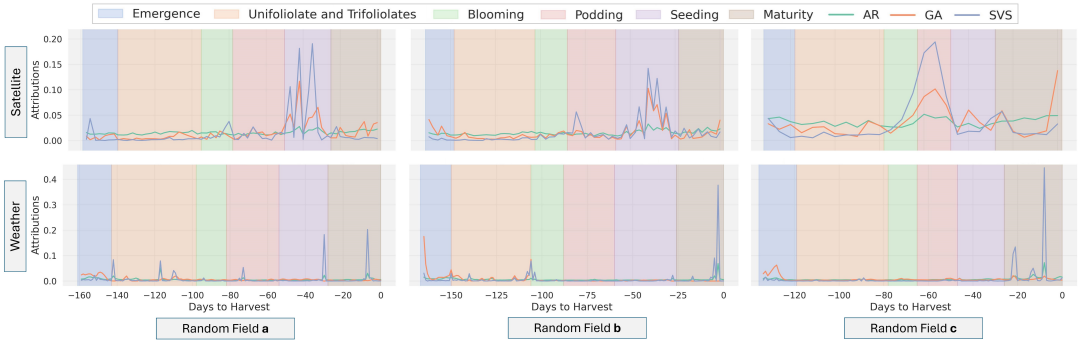


Fig. 3.31.: Field-level average attributions of the satellite and weather modalities, for three random soybean fields. The six growth stage periods are shown in the background of each plot.

Infidelity scores

Sensitivity scores

c. Quantitative Evaluation The sensitivity and infidelity metrics are used to evaluate and compare the robustness of the attributions generated by the three methods. Each metric assigns a single-valued score per pixel, with smaller values indicating greater robustness and stability. Figure 3.30 presents these results across all corn fields, using 32 pixels per field. We observe that the infidelity scores are consistently low across all three methods, indicating a strong alignment between the magnitude of the attributions and the impact that input perturbations have on the model's predictions. It further suggests that all attribution methods effectively capture the relationship between input features and model outputs. In contrast, the sensitivity scores are notably higher, particularly for the attributions generated by the GA method. This result corresponds to the findings in subsection 3.3.4, where GA was observed to provide inconsistent and distant attributions for pixels within the same field. The stability of the AR attributions is observed across other crops and regions, as we demonstrate in Appendix E.2.

Growth stage attributions

d. Agronomic interpretation The attribution results can potentially offer valuable agronomic insights about the growth stages most influential for the model's predictions. Such analysis requires additional information about the start and end dates of the crop's growth stages at each field. While this metadata was not consistently measured throughout the growing season, we obtained approximate growth stage information for a subset of soybean fields[4]. Figure 3.31 illustrates three examples of soybean fields, where the temporal attributions are presented alongside the corresponding growth stages. To address **RQ4**, we study the alignment between the identified critical growth stages and established agronomic knowledge, similarly to the analysis conducted in 3.2.5. A detailed discussion is included in Appendix E.2.

Weather Events

Connecting attributions with weather events

To investigate the possible impact of special weather events on their

[4]Phenology data was provided by www.xarvio.com, using their in-house developed and commercially deployed growth stage models.

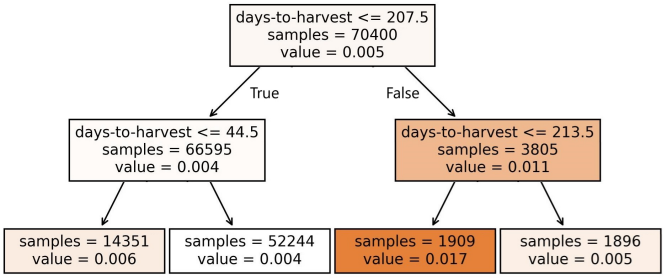


Fig. 3.32.: Decision Tree with two levels, predicting the AR temporal attributions of the weather Transformer encoder. The color of each box is used as a scale for the predicted attribution values.

attribution score, we train a decision tree model to predict the attribution of each time step based on its weather properties: minimum, average, and maximum daily temperatures, as well as total precipitation. We additionally include the number of days before harvest among predictive features, allowing the model to contextualize each weather event within the growth cycle of the crop. Specifically, we randomly sample 200 pixels from each field, merge the associated weather time series together, shuffle the instances to break the sequences, and then partition the datasets into 80% for training and 20% for testing. We train a separate decision tree for each set of corn fields belonging to the same farm and the same year. We experiment with decision tree depths of two and three, to ensure the learned models remain interpretable.

Experimental setup

Figure 3.32 presents the results of a two-level decision tree model trained on data from a farm with two fields from the 2023 season, using AR scores as the target attributions. The R^2 scores of the decision tree for this farm were particularly high and thus reliable for interpretation, reaching 0.75 in both the training and test sets. The figure represents the splitting of the training set, composed of 70,400 samples. We observe that the number of days before harvesting is the only variable used to split the tree. The darkest leaf in the tree, representing 2.7% of the training set (1909 samples), shows a notably high attribution score of 0.017. These high-importance events occur between 213 and 207 days before harvesting. The remaining instances have attributions between 0.004 and 0.006, covering 97.3% of the training samples. These results reveal where the highest mass of the attributions is located within the growth cycle for the considered farm.

Results of a bilayered decision tree

Influential weather events

A slight increase in the tree depth can improve the tree performance across multiple farms while maintaining interpretability. Figure 3.33 illustrates the weather events decision tree, of three levels, for a farm of three fields from the year 2023. For this farm, the tree model uses weather variables for splitting, in addition to the number of days before harvesting of each instance. It achieves an accuracy of 79% on both the training and test sets, on the task of predicting the AR temporal attributions. We observe that the left branch of the tree covers a large portion of the training samples, greater than 95.5%. This branch includes 90% of the instances with attribution scores ranging from 0.004 to 0.005, corresponding to days between 10 and 199 prior to harvesting. To identify

Results of a three-layered decision tree

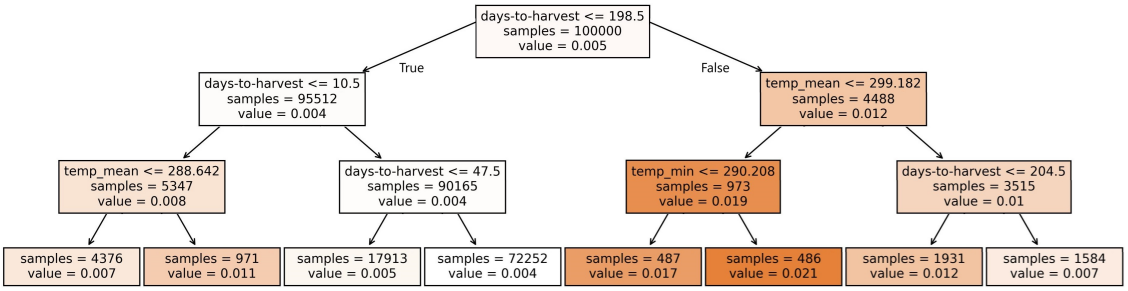


Fig. 3.33.: Decision Tree with three levels. The results shown are on the train set of 3 fields from the same farm, from 2023, predicting the AR temporal attributions of the weather Transformer encoder.

Influential weather events

the samples most critical to the model's predictions, we focus on the nodes and leaves with the darkest shading. Approximately 1% of the training samples occurring earlier than 198 days before harvesting, and associated with a mean temperature below 299.2 K, exhibited attribution scores between 0.017 and 0.021. Particularly, days where the minimum temperature exceeded 290 K attained the highest attribution score of 0.021. This finding indicates that such weather events are highly influential in the Transformer model, suggesting a critical role that specific temperature conditions play in the early days of the growing cycle.

Modality Importance

After evaluating temporal attributions for satellite and weather modalities in Section 3.3.5 and analyzing their role in identifying important weather events in Section 3.3.5, we now address **RQ5** and investigate two modality attribution methods in order to compare the relative impact of the different modalities on model predictions. The first method is derived from SVS scores, while the second one is a newly proposed method, Weighted Modality Activation (WMA), based on the inner parameters of the model, as we describe in the following.

a. SVS-based modality attribution SVS can estimate the contribution of each individual input feature to the model's predictions. To get the relative importance of different data modalities, we aggregate the absolute SVS scores per modality. Specifically, for each pixel, we compute the importance score of each input feature by taking the absolute values of the SVS scores, which are then summed separately for each modality. To ensure comparability, we subsequently scale the modality scores such that they sum to one.

b. Weighted Modality Activation (WMA) Since the multimodal networks described in Section 3.3.3 use a concatenation-based fusion followed by a linear layer, we propose to exploit this structure to infer modality attributions. We can rewrite the final prediction $\hat{y}_i$ of sample i as the

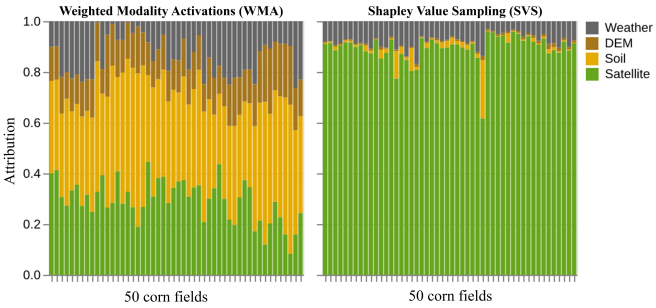


Fig. 3.34.: Comparing the modality scores for the same random set of 50 corn fields.

weighted combination of the modality activations $\mathbf{z}_i = \text{concat}\left(\mathbf{z}_i^m\right)$,, with $m \in \{$ satellite (sa), weather (w), soil (so), dem $\}$:

$$\hat{y}_i = \mathbf{w}.\mathbf{z}_i + b = \sum_m \mathbf{w}^m.\mathbf{z}_i^m + b = \sum_m \hat{y}_i^m + b,$$

and infer modality relevance scores $\mathcal{R}_i^m$:

$$\mathcal{R}_i^m = \frac{\hat{y}_i^m}{\hat{y}_i - b},$$

where $\mathbf{w} = \text{concat}\left(\mathbf{w}^{sa}, \mathbf{w}^{w}, \mathbf{w}^{so}, \mathbf{w}^{dem}\right)$ and b are the weights vector and bias of the final regression layer, respectively. This approach can be viewed as an alternative to CAM and Grad-CAM methods [348, 267], which are widely used for explaining classification tasks in computer vision. However, while CAM and Grad-CAM are specifically designed for convolutional networks operating on a single modality, our method is applicable to any multimodal regression task using a concatenation fusion mechanism and a MLP as a regression head. Furthermore, it can be extended to various differentiable fusion strategies and regression heads using gradient-based techniques, similar to how Grad-CAM extends CAM.

Similarity between WMA and CAM

c. Results comparison We compute the SVS and WMA modality scores on a random selection of 32 pixels per field, using the same pixels sampled in Section 3.3.5. The scores are then aggregated per field by averaging the modality scores across the 32 pixels. In Figure 3.34, we compare both methods and present the modality scores for 50 corn fields.

WMA results

WMA scores indicate that soil features have the highest impact on the prediction, accounting for an average of 41.3% across all fields, followed by satellite data at 29.4%. Terrain elevation features and weather data have the smallest share of contributions, with an average importance of 15.1% and 14.2%, respectively. In contrast, Shapley values indicate a different distribution of relative importance, with satellite data contributing the predominant share at 89.5% on average, followed by weather at 7.9%. Soil and DEM features have only a minimal impact, contributing less than 2% and 1%, respectively.

SVS results

Extended results

We attach in Appendix E.3 the results of the same comparison for other

crops and regions, in which the satellite modality remains predominant according to the SVS results, while it contributes much less according to the WMA. This difference can be particularly attributed to the computational process: WMA relies only on the regression head to infer modality scores, while the SVS method uses the entire model.

Related work

Notably, a similar study on yield prediction computed modality scores resembling our WMA attributions [213]. Although their results align more closely with SVS scores, their model incorporates two key modifications: (1) replacing our Transformer encoders with LSTMs, and (2) employing an attention-based fusion approach rather than simple concatenation. These architectural differences prevent a direct comparison with our results.

3.3.6 Summary

We summarize in the following the main takeaways of this chapter. We recall the research questions we defined and provide their responses in the light of the experimental results discussed above.

RQ1: Why was the Transformer-based model chosen?

To process multimodal data, we designed networks with an intermediate fusion mechanism, enabling the training of modality-specific encoders to address the unique characteristics of each modality effectively. We evaluated convolutional, recurrent, and Transformer-based networks, comparing their accuracy and inference times. The results demonstrated that the Transformer-based model offers a good balance between high performance and inference speed, in addition to its potential to provide intrinsic interpretability of its predictions.

RQ2: What did the analysis of the intermediate representations reveal?

The linear probing experiments revealed that satellite representations exhibit a significantly stronger linear correlation with the predictions compared to representations from other modalities. This linearity improves progressively across deeper layers of the model. Further analysis of the learned attention weights within the satellite encoder showed that the first three Transformer layers generate similar weights for pixels within the same fields, while the fourth layer introduces greater diversity at the field level. Importantly, these variations in attention weights were found to be uncorrelated with differences in predicted yield. Examining the temporal distribution of attention weights uncovered distinct patterns of key time periods for the satellite and weather encoders. Entropy analysis highlighted that important weather information is concentrated within a few critical time steps, while the satellite information relevant for predictions is distributed more evenly across the entire growth cycle.

RQ3: Which method for estimating temporal attributions is most reliable?

We compared two model-specific methods, namely AR and GA, against a model-agnostic technique, SVS. Due to the impracticality, or even un-

feasibility, of acquiring ground truth labels for the feature importance scores at each pixel, we relied on comparative analyses and quantitative evaluation metrics to assess the attribution methods. The infidelity and sensitivity scores highlighted that AR delivers more consistent and robust attributions compared to the GA method. While SVS demonstrated performance close to AR, we would favor the latter for two primary reasons: First, the intrinsic nature of AR enhances its faithfulness to the model, ensuring the attributions are more aligned with the model's internal mechanisms. Second, AR has significantly faster computation times compared to the lengthy calculations and perturbations required by the Shapley-based method. These factors collectively make AR a preferable choice for interpreting model predictions in this context.

RQ4: Can the temporal attributions provide agronomically relevant insights?
We acquired crop phenology information for certain soybean fields in Argentina and demonstrated how to interpret temporal attributions in the context of agronomic knowledge. As detailed in E.2, the availability of starting and ending dates for each growth stage in each field allowed us to validate the model's reasoning against established expert insights. Specifically, we examined three soybean fields and showed how certain growth stages known to be critical for the soybean yield were also important for the model, and explained that certain visual indicators of the yield at certain stages might also have been used by the model for its predictions. In contrast, some critical growth stages appeared to have little influence on the model's decisions. This raises questions about potential gaps in the model's reasoning, calling for further experiments across multiple fields to validate these observations and explore their implications on the model performance.

What is the utility of weather events' analysis?
Temporal attributions of the weather data highlight the significance of specific time periods for the model's predictions. The analysis of weather events conducted in subsection 3.3.5 explores whether particular weather events have a substantial influence on the model's decisions. When using a two-level decision tree, the findings suggest that attribution scores are more dependent on temporal factors than on climate conditions. In contrast, the three-level tree-based model incorporate temperature levels among its key splitting criteria, revealing the correlation between highly important time steps and their weather properties.

RQ5: Which modality importance estimation method is most reliable?
Shapley values stand out for their ability to capture feature interactions by employing principles from game theory, considering multiple feature subsets and their contributions to the model before inferring feature attributions. Notably, SVS modality importance results exhibit a stronger alignment with the linear separability observed in the linear probing analysis, which indicated that satellite representations were most corre-

lated with the final prediction. In contrast, the strength of WMA scores lies in their inherent connection to the model's architecture, which makes their importance estimations more faithful to the model's behavior [258]. Evaluating the correctness of these methods remains challenging, as the modality impact scores do not necessarily reflect the agronomic significance of each modality, where established field knowledge could have been leveraged as a reference. Instead, these scores indicate how the model uses each modality, which depends on its learning scheme and the data patterns it captured during training.

Overall, this work highlights the potential of leveraging intrinsic interpretability within Transformer-based models to enhance understanding in multimodal learning frameworks. We examined the learned representations for each modality, inferred temporal attributions using both model-specific and model-agnostic approaches, and proposed WMA, an intrinsic method to derive modality importance scores. Our experiments revealed the AR as a reliable intrinsic method for estimating temporal attributions, outperforming both SVS and GA methods. In contrast, the modality contributions evaluated by SVS indicated that satellite data has a predominantly high influence on the predictions, whereas WMA suggests a more evenly distributed contribution across the four input modalities.

Limitations A notable limitation of this study arises from the variability in seeding and harvesting dates across different fields. This variability complicates the comparison of temporal attribution results at the dataset level, as the sequence lengths of the temporal modalities varies between fields. Adding to this challenge is the missing phenology information of various growth stages for most fields. Furthermore, the modality attribution analysis did not yield relevant insights due to the conflicting results obtained between WMA and SVS estimations.

Future work Follow-up studies should prioritize a detailed analysis of the modality attribution methods to explain the conflicting results, and a quantitative evaluation to determine the most reliable approach for assessing the relative importance of different modalities in yield prediction tasks. Additionally, obtaining detailed growth stage data is essential for extending agronomical analyses across multiple fields and deriving generalizable insights. Ultimately, resolving challenges related to the interpretability can facilitate building on the explainability findings to enforce certain rules or constraints during the learning phase, potentially optimizing the model performance.

4 Multi-Task Learning

4.1 INTRODUCTION

In prior sections, we discussed how the multimodal setup calls for advanced modeling techniques, often leading to increased model complexity, which comes at the expense of model interpretability [154, 117]. After having explored various explainability methods for such networks in case of early and intermediate modality fusion in Chapter 3, we investigate in this chapter a different approach to explaining model predictions in the context of multimodal learning: we explore how modalities can be leveraged through multitask learning to intrinsically interpret model predictions. In particular, instead of additional inputs, we use certain modalities as additional targets to be predicted along with the main task. We show how this modeling context provides numerous benefits: (1) the model performance remains comparable to the multimodal baseline performance, and in some cases achieves better scores, (2) prediction errors in the main task can be explained via the model behavior in the auxiliary task(s), (3) in case of data scarcity, the additional modalities do not need to be collected for model inference at deployment. We demonstrate our approach on three remote sensing datasets, including segmentation, classification, and regression tasks.

Benefits of switching from multimodal to multitask learning

4.2 RELATED WORK

4.2.1 Multimodal and multitask learning

Combining multiple modalities into a common pipeline is a common practice that aims at improving model performance, whenever suitable diverse data is available. In fact, it was shown that models fusing data from different modalities outperform their uni-modal counterparts both intuitively and provably [138]. Similarly, multitask learning is leveraged to predict multiple targets using a shared model, achieving in most cases smaller memory footprint, reduced number of calculations, and improved performance [203, 63, 198, 160, 268, 190, 179, 344, 314]. There are still certain scenarios in which single task networks might outperform multitask counterparts, due to the number of tasks, their types, and the accuracy of their annotated labels [314, 284, 231, 299]. Standley et al. [284] argue that this improvement is not guaranteed, as multitask learning can sometimes degrade model performance. They attribute this to multiple factors, including the varying learning rates required for different tasks, the dominance of one task over others, and gradient interference, which complicates the optimization process.

Benefits of multitask learning

Limitations of multitask learning

4.2.2 *Explainability through multitask learning*

Joint training One of the few intrinsic methods in XAI that relies on multitask learning is **joint training**. This method generates explanations by augmenting the original network with additional tasks to explicitly return textual, imagery or numerical explanations, along with the model decisions [127, 241, 189, 158, 254, 192, 295]. Hendricks et al. [127] apply joint training to explain an image classification task of bird species, by predicting the class label along with a textual explanation. Their proposed method relies on sampling techniques and reinforcement learning, to ensure the explanation describes visual content present in the input image and contains appropriate information related to the predicted class. Similarly, Liu et al. [189] propose a framework to generate textual explanation, among other types of fine-grained explanations, for classification tasks in natural language processing. In another language-based application, Tang and Surdeanu [295] propose an approach for relation extraction that jointly learns how to explain and predict. Their approach consists of learning two tasks: the first one is a relation classification which predicts the relation that holds between two given entities, while the second task is an explainability classification which labels words in the textual context where the extracted relation is expressed as important or not.

Prediction of a textual explanation

Prediction of a binary label as explanation

Semantic bottlenecks Another line of explanation methods close to the joint training family are **semantic bottleneck networks**. Such models were introduced by Losch et al. [194] and consist of defining an intermediate *bottleneck* layer where features are enforced to align with semantic concepts. The proposed encoder-decoder design has the downside of non-linearity between the representations learned at the bottleneck layer and the final predictions. To address this issue and improve model interpretability, Marcos et al. [204] proposed to place the semantic layer right before the final layer, enabling a linear mapping between the concepts and the predictions. This approach was applied in different applications. For instance, Levering et al. [179] predicts a landscape scenicness score over satellite images, and use land cover classes as an interpretable intermediate task to scenicness regression. Mojab et al. [219] predict glaucoma, one of the leading causes of blindness worldwide, by generating two segmentation masks to locate the optic disc and optic cup regions in a fundus image, before fusing the maps together to assess the cup-to-disc ratio and predict the presence of glaucoma. Echterhoff et al. [73] predict user and vehicle behavior in the context of human-assisted or autonomous driving. Their proposed model includes a concept bottleneck based on visual features to predict and explain driving control commands.

Encoder-decoder design for semantic bottlenecks

Enhancing linearity between concept and prediction spaces

Our approach While previous methods provide explicit explanations for model predictions, this is limited in some cases by the availability of training labels for the explanation task, in the form of semantic labels for the semantic bottleneck approach, or explicit sentences and scores for the joint training framework. In our work, we overcome this limitation

Usage of available modalities as explanatory tasks

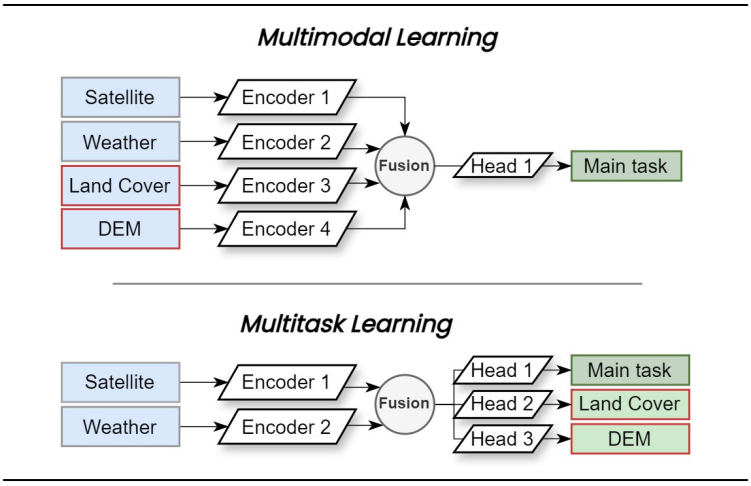


Fig. 4.1.: Comparison of multimodal against multitask setups in a RS dataset. DEM refers to digital elevation maps.

by relying on available input modalities and turning them into auxiliary tasks. While this method does not provide explicit explanations, we explore how to extract insightful results from this framework to intrinsically explain model predictions through the auxiliary task(s). In Section 4.3, we compare model performance between multimodal and multitask setups. Subsequently, in Section 4.4, we examine the interpretability of the multitask experiments.

4.3 MODELING AND PERFORMANCE

In multimodal datasets, additional modalities are typically incorporated as input data. Yet, not all of them are essential for achieving the baseline model performance. In particular, satellite imagery in RS datasets inherently encodes a rich and diverse range of information about the Earth's surface. For instance, multispectral sensors capture spectral characteristics across multiple bands, while SAR sensors provide structural and textural details. Exploiting satellite data characteristic, we focus on RS multimodal datasets and explore in this section the effect of shifting additional modalities between input data and auxiliary tasks, as depicted in Figure 4.1. To maintain a robust baseline, we consistently incorporate satellite data as one of the input modalities across all multitask experiments, to avoid significant performance degradation.

To validate our approach, we compare model performance under both the multimodal and multitask setups. We use three RS multimodal datasets, covering segmentation, classification, and regression tasks. In the following subsections, we provide a description of these datasets, outline the experimental setup for multimodal and multitask modeling, and present the results for each dataset.

4.3.1 Datasets

CropYield for Yield Prediction The *CropYield* dataset used in this chapter is the same as that presented in the previous one, but only data from Argentina is considered (see Section 3.3.2). It contains approximately 500 crop yield maps of corn, soybean, and wheat fields, covering crop seasons from 2017 to 2023. Since the dataset is processed on a pixel-wise basis, it counts more than 3.5 million input samples. The modalities we use in this chapter are the satellite multispectral imagery, weather data, DEM properties, and crop type. The crop type modality is newly introduced in this study, as we merge samples from different crops, whereas in the previous chapter, each crop type and region was processed using separate models. Both satellite and weather data are temporal, spanning from seeding to harvesting dates each year. Yield maps rasterized at a 10-meter resolution are used as the main target values. Since this dataset is not publicly released, we further validate our approach using two additional, publicly available, RS datasets: Benge and TreeSAT .

Benge for Land Cover Segmentation *Benge* is an open-source[1] multimodal dataset for Land Use and Land Cover (LULC) segmentation, extending the BigEarthNet dataset [220, 289, 290]. BigEarthNet contains SAR and multispectral satellite images, from Sentinel-1 (S1) and S2 missions respectively, for 590,326 locations throughout Europe. Benge complements this dataset by providing additional data for each of these observations, including elevation maps, weather data, climate zone information, and seasonal encoding. For weather data, we include all five weather features (temperature, two wind vectors, relative humidity, and atmospheric pressure) when the modality is used as input data. In contrast, when using weather data as an auxiliary target, we exclude wind vectors. Following the recommendations of [220], experiments were initially conducted on a small subset of the dataset. Subsequently, the best-performing architectures were trained on the 0.2 split of the full dataset, in order to balance computational efficiency with comparable performance.

TreeSAT for Tree Identification *TreeSAT* is an open-source[2] dataset for tree species classification in Central Europe based on multi-sensor data from aerial imagery and satellite observations, including SAR (S1) and multispectral (S2) satellite images [5]. The dataset consists of 50 381 image patches and contains labels of 15 tree genera (the main classification task), nine forest stand types, and three foliage types, corresponding to classification levels 3, 2, and 1, respectively. These classification levels are referred to as L3, L2, and L1. Additionally, it includes an approximation of tree age, which is treated as a continuous rather than a categorical feature. The labels are derived from the forest administration data of the federal state of Lower Saxony, Germany.

[1] https://github.com/HSG-AIML/ben-ge
[2] https://zenodo.org/records/6780578

Tab. 4.1.: Available modalities, encoders, prediction heads, loss functions and evaluation metrics used in the three multimodal datasets: CropYield , Benge , and TreeSAT . The main input and target modalities are highlighted in bold.

Dataset	Modality	Type	Encoder	Prediction Head	Loss function	Metric
CropYield	**Sat (S2)**	TS of 25 features	Transformer	-	-	-
	Weather	TS of 4 features	Transformer	-	-	-
	Yield	single scalar	-	Reg. MLP	MSE	R^2
	Crop label	3 classes	MLP	Class. MLP	Cross entropy	micro-F1
	DEM	5 features	MLP	Reg. MLP	MSE	MAE
Benge	**Sat (S1,S2)**	multichannel image	U-Net	-	-	-
	LULC	segmentation mask	-	Multiclass Segmentation	Cross entropy	IoU
	Elevation	single channel image	U-Net	Dense Segmentation	MSE	MAE
	Climate Zone	12 classes	Embeddings	Class. MLP	Cross entropy	micro-F1
	Season	single scalar	MLP	Reg. MLP	MSE	MAE
	Weather	5 features	MLP	Reg. MLP	MSE	MAE
TreeSAT	**Aerial**	multichannel image	CNN	-	-	-
	Sat (S1,S2)	multichannel image	MLP	-	-	-
	Level-3 (L3)	15 classes	-	Class. MLP	Cross entropy	micro-F1
	Level-2 (L2)	9 classes	-	Class. MLP	Cross entropy	micro-F1
	Level-1 (L1)	3 classes	-	Class. MLP	Cross entropy	micro-F1
	Age	single scalar	MLP	Reg. MLP	MSE	MAE

Reg.: Regression | TS: Time Series | Sat: Satellite | S1: Sentinel-1 | S2: Sentinel-2

Table 4.3.1 provides a summary of the modalities used in each dataset, highlighting the main input modalities and the main target.

4.3.2 Experimental Setup

Modality Encoders We evaluate and compare the performance of both multimodal and multitask learning on multimodal datasets. Given the diversity of the input data types, we adopt an intermediate fusion approach: each input modality is processed by a dedicated encoder, generating an intermediate representation, which is then fused across modalities before being passed to a task-specific head for the final predictions. The intermediate fusion approach is flexible in handling multiple input modalities despite differences in data type, spatial characteristics, or temporal resolution. This approach has also often outperformed early and late fusion techniques, particularly in RS applications [212]. The architecture of the encoder is chosen based on the types of the input and the target. For *imagery inputs*, we either use a U-Net architecture in segmentation tasks or a convolutional network in other tasks. If the input image is small, such as in low-resolution satellite imagery, we flatten it and process it using a MLP. *Time-series inputs* are processed using Transformers, including positional encoding based on each timestamp. *Tabular data* are processed using MLPs, whether they include a single or multiple features. Finally, for *categorical inputs*, we use a MLP or an embedding layer.

Intermediate fusion pipeline

Encoder types for each input type

Intermediate Fusion The intermediate representations generated by the modality encoders are combined at the fusion block through concatenation, optionally followed by convolutional layers. For *regression* and *classification tasks*, each encoder outputs a 1-dimensional feature vector representing its respective modality. These vectors are simply concatenated at the fusion stage, with no additional processing. For *segmentation tasks*, modalities are encoded into a three-dimensional latent representation (channels $\times$ height $\times$ width). If the input is an image processed via a U-Net, this representation is obtained naturally. For tabular data encoded through a MLP, the one-dimensional output can be expanded into additional dimensions to align with the spatial structure of other representations. This alignment facilitates the concatenation along the channel dimension, followed by additional convolutional layers that preserve the spatial dimensions (height and width) of the fused representation.

Intermediate fusion mechanism per task

Prediction Heads Multiple prediction heads can branch out from the fusion block, each dedicated to a specific target. For *segmentation tasks*, the prediction head consists of convolutional layers, in order to preserve the spatial dimensions of the image. For *regression and classification tasks*, a MLP is used to return the appropriate number of output neurons for the task.

Prediction head per task

Loss and Metrics The optimization loss for each task is defined based on its nature. For *classification tasks*, including semantic segmentation, the cross-entropy loss is used, whereas *regression tasks*, including dense segmentation, we use the mean squared error (MSE) function. In the

Loss function per task

multitask learning scenario, the loss contributions of individual tasks are manually fine-tuned. For example, we evaluated strategies such as equally distributing the loss contribution across all tasks, or prioritizing the primary task by assigning it a higher weight (e.g., 60% or 80%) while maintaining a uniform distribution of weights for the auxiliary tasks. To further evaluate and report performance, additional metrics are included. MAE and R^2 are used for regression and dense segmentation tasks, the F1 score for classification tasks, and the intersection over union (IoU) for semantic segmentation tasks.

Loss weighting in multitask learning

Evaluation metrics

In Table 4.3.1, we provide a summary of the encoder, prediction head, loss function, and evaluation metric used for each modality in each dataset.

4.3.3 Results

In this section, we analyze the performance results of the different modeling setups, including baselines, which include the remotely sensed images (aerial and satellites) and temporal modalities, multimodal learning experiments, which test different combinations of additional input modalities, and multitask learning experiments, which shifts some modalities from being additional inputs to auxiliary targets.

CropYield Table 4.2 combines the results of multiple CropYield experiments. The first three experiments train the model using satellite data alone, based on the subset data of each crop individually, while all subsequent experiments merge samples from all crop types. The first four baseline experiments indicate that combining crop types has a positive impact on the overall model performance, achieving the relatively high R^2 score of 0.81. Evaluating the performance per crop type reveals an increase of 0.15 and 0.04 in the R^2 score of wheat and corn pixels, respectively. Nevertheless, a notable decline of 0.19 is observed for soybean fields. Despite using weighted data sampling during the training to mitigate class imbalance, these results correlate with the size of each crop type within the dataset, as we observe that the smallest crop subset (wheat) benefits the most, followed by the second smallest (corn). In contrast, the largest dataset (soybean) exhibits a decline, and performed better when trained individually, in Experiment **1**. As a result, corn and wheat samples benefit from the data mixing, unlike soybean samples.

Impact of mixing crops during training

In multimodal setups (Experiments **5**-**9**), a performance comparable to the baseline experiment **4** is observed when including weather and DEM as additional inputs to the model, in Experiment **8**, while any other combination of auxiliary inputs yields a decline in the performance. Surprisingly, this includes Experiments **5**, **7**, and **9**, where we provide the model with the crop label of each pixel sample. In contrast, in multitask setups (Experiments **10**-**13**), we observe that forcing the model to predict the crop label improved its performance, particularly when including weather and DEM modalities as inputs, in Experiment **12**, and when including no additional input modality, in Experiment **10**. The latter

Comparing multimodal against baseline experiments

Comparing multitask against multimodal experiments

Tab. 4.2.: Modeling performance on the test set of the CropYield dataset. The best and second-best scores are highlighted in bold and underlined, respectively. Crop classification performance is given in micro F1 score.

Experiment		Modalities				Main task				Auxiliary tasks	
		Sat	Crop label	Weather	DEM	Yield (R^2)	Yield (R^2-Soybean)	Yield (R^2-Wheat)	Yield (R^2-Corn)	Crop cls. (F1)	DEM (MAE)
Baselines	1	→□	(soybean)			0.64	**0.64**	-	-	-	-
	2	→□	(wheat)			0.64	-	0.64	-	-	-
	3	→□	(corn)			0.48	-	-	0.48	-	-
	4	→□	(all crops)			0.81	0.45	0.79	0.62	-	-
MML	5	→□	→□			0.77	0.44	0.78	0.51	-	-
	6	→□		→□		0.75	0.37	0.80	0.45	-	-
	7	→□	→□	→□		0.79	0.40	0.78	0.59	-	-
	8	→□		→□	→□	0.81	0.45	0.78	**0.63**	-	-
	9	→□	→□	→□	→□	0.79	0.42	0.75	0.57	-	-
MTL	10	→□	□→			**0.82**	0.52	**0.82**	**0.63**	99.4	-
	11	→□	□→	→□		0.77	0.48	0.77	0.49	**99.5**	-
	12	→□	□→	→□	→□	0.80	0.43	0.75	0.60	**99.5**	-
	13	→□	□→	→□	□→	0.75	0.37	0.78	0.48	99.3	**0.42**

→□ Input | □→ Output | MML: Multimodal learning | MTL: Multitask Learning | Sat: Satellite | cls.: classification.

even reached the highest overall R^2 score across all experiments, along with the best score in wheat and corn samples, and the second best score for soybean samples. The model further reached a very high F1-score of 99.4% in the crop classification task, which brings a great benefit in practice, enabling the distinction of crop types along the accurate yield prediction.

Influence of the crop label modality

We assume that the performance gap in yield prediction between Experiments **5** and **10** is due to the shared representation of the multitask learning setup, in which the model is forced to learn representations related to the different crop labels, which positively influences the accuracy of the predicted yield. Moreover, the gap observed between the global vs. crop-specific R^2 scores is caused by the nature of this score, and the gap confirms that the model's performance is not consistent across different crop types.

In the explainability analysis in Section 4.4.1, we will focus on Experiment **10**, which predicts the yield and crop labels using the satellite data alone.

Benge We present the results of Benge dataset in Table 4.3. In the baseline experiment, the model is trained on the multispectral and SAR satellite images alone, achieving the second best scores in the main task of LULC, with an accuracy of 87.94% and an IoU score of 0.388. In the multimodal experiments (**2**-**8**), we evaluate different combinations of one or more additional input modalities, prioritizing elevation data due to its spatial dimension, which the remaining modalities lack. While all multimodal experiments yielded results comparable to the baseline, Experiment **7** including the elevation and weather data have slightly outperformed it, achieving an accuracy of 87.95%. Similarly, Experiment **4**, which includes seasonal information, achieves a marginally higher IoU score of 0.389, also surpassing the baseline.

Baseline performance

Multimodal experiments

In the multitask setup (Experiments **9**-**15**), the LULC accuracies remain within a range similar to the baseline and multimodal experiments, while IoU scores marginally declined. Notably, certain modality combinations reached improved accuracies when incorporated as auxiliary tasks rather than as input modalities, such as the climate zone (in Experiments **3** and **10**) and the combination of all modalities (in Experiments **8** and **15**). Regarding the performance of the auxiliary tasks, we observe that climate zone classification (with 12 classes) achieves a high F1 score close to 95%. Similarly, the season prediction task yields very low MAE scores, particularly in comparison to the errors observed in elevation and weather predictions. It is important to note that weather and season data are normalized, whereas elevation values range between 0 and 1.

Multitask experiments

Performance in auxiliary tasks

Overall, we find that the additional input modalities in the multimodal setup do not contribute to improved model performance. However, our results remain consistent with the scores reported in [220]. In contrast, the multitask setup neither degrades nor enhances the primary task's performance, but its other benefits persist. In Section 4.4.2, we further investigate the explanatory capacity of each output modality, using Experiment **15** as a testbed.

Comparing multimodal against multitask setups

Tab. 4.3.: Test set performance on the Benge dataset. The best and second-best scores are highlighted in bold and underlined, respectively. Climate zone classification performance is given in micro F1 score.

Experiment		Modalities					Main task		Auxiliary tasks			
		Sat	Elevation	Climate Zone	Season	Weather	LULC (Accuracy)	LULC (IoU)	Elevation (MAE)	Climate zone (F1)	Season (MAE)	Weather (MAE)
Baseline	1	→□					<u>87.94</u>	<u>0.388</u>	-	-	-	-
MML	2	→□	→□				87.91	0.386	-	-	-	-
	3	→□		→□			87.90	0.386	-	-	-	-
	4	→□			→□		87.91	**0.389**	-	-	-	-
	5	→□				→□	87.93	0.387	-	-	-	-
	6	→□	→□		→□		87.90	0.385	-	-	-	-
	7	→□	→□			→□	**87.95**	0.383	-	-	-	-
	8	→□	→□	→□	→□	→□	87.85	0.387	-	-	-	-
MTL	9	→□	□→				87.90	0.380	<u>0.162</u>	-	-	-
	10	→□		□→			87.93	0.379	-	**94.88**	-	-
	11	→□			□→		87.91	0.380	-	-	**7e-8**	-
	12	→□				□→	87.91	0.381	-	-	-	<u>0.018</u>
	13	→□	□→		□→		87.91	0.377	<u>0.162</u>	-	<u>9e-8</u>	-
	14	→□	□→			□→	87.89	0.377	**0.161**	-	-	<u>0.018</u>
	15	→□	□→	□→	□→	□→	87.89	0.373	<u>0.162</u>	<u>94.77</u>	5e-5	**0.015**

→□ Input | □→ Output | MML: Multimodal Learning | MTL: Multitask Learning | Sat: Satellite images.

Tab. 4.4.: Test set performance on the TreeSAT dataset. The best and second-best scores are highlighted in bold and underlined, respectively. *Images* refer to the aerial and two satellite images (from Sentinel-1 and Sentinel-2 missions). L3, L2, and L1 classification performance are given in micro F1 score.

Experiment		Modalities				Main task	Auxiliary tasks		
		Images	L2	L1	Age	L3 (F1)	L2 (F1)	L1 (F1)	Age (MAE)
Baseline	1	→□				<u>74.3</u>	-	-	-
MML	2	→□			→□	**76.9**	-	-	-
MTL	3	→□	□→			<u>74.3</u>	**78.2**	-	-
	4	→□		□→		70.3	-	<u>92.1</u>	-
	5	→□			□→	71.8	-	-	**0.52**
	6	→□	□→	□→		71.1	76.6	**92.3**	-
	7	→□	□→		□→	72.2	<u>77.3</u>	-	**0.52**
	8	→□	□→	□→	□→	70.4	75.5	<u>92.2</u>	<u>0.53</u>

→□ Input | □→ Output | MML: Multimodal Learning | MTL: Multitask Learning.

TreeSAT Compared to CropYield and Benge , TreeSAT dataset experiments exhibit different patterns, as shown in the results displayed in Table 4.4. The baseline model, trained on the three imagery modalities (i.e. aerial imagery and two satellite images), achieves a micro F1-score of 74.3%, ranking second. This represents a significant improvement compared to the 71.66% accuracy reported by Ahlswede et al. [5], despite also using their best-performing model architecture. As shown in Table 4.4, the highest accuracy of 76.9% is reached by the multimodal experiment that includes the age as an additional input data. Tree type labels from levels 1 and 2 were not included as input features, as acquiring this data at inference time would be impractical in real-world scenarios. In contrast, age can, in some cases, be inferred from historical records and old maps, which document events such as deforestation, wildfires, or planting.

Baseline performance

Multimodal setup

Multitask setup

In the multitask experiments, the primary task's performance declines slightly compared to the multimodal experiment, but maintains F1-scores above 70%. Specifically, Experiment **4**, which includes only the first level (L1), and Experiment **8**, which infers all modalities, yield the lowest L1 F1-scores of 70.3% and 70.4%, respectively. In contrast, including the second level (L2) in Experiment **3** achieved the same accuracy as the baseline model (74.3%) while also yielding accurate labels for the second level labels, reaching a micro F1-score of 78.2%.

Performance in auxiliary tasks

Overall, in the multitask experiments, L2 classification (with 9 classes) demonstrates high accuracy, L1 classification (with 3 classes) achieves significantly better scores, while age prediction (with normalized values) exhibits moderate performance. Experiment **7**, which reached the second performance in the main task among multitask experiments, will be explored in the explanatory analysis in Section 4.4.3.

4.4 INTERPRETABILITY THROUGH AUXILIARY TASKS

We propose in this section leveraging multimodal datasets in RS to enhance model interpretability via the multitask learning framework. Even when improvements in the main task's performance are not guaranteed, a multitask learning setup based on hard-parameter sharing facilitates the learning of a shared intermediate representation for multiple downstream tasks. This setup has the potential to enhance interpretability across tasks. Specifically, we demonstrate how correlated tasks influence each other during training, whether positively, where accuracy improvements in one task benefit another, or negatively, through error propagation.

The results in the previous section have shown that the multitask models achieved comparable performance the the baseline and multimodal experiments. Subsequently, for each dataset, we pick an example from the multitask experiments, and demonstrate in the following how to conduct the interpretability analysis. Our focus will be on explaining errors in the main task in light of the predictions from the auxiliary task(s).

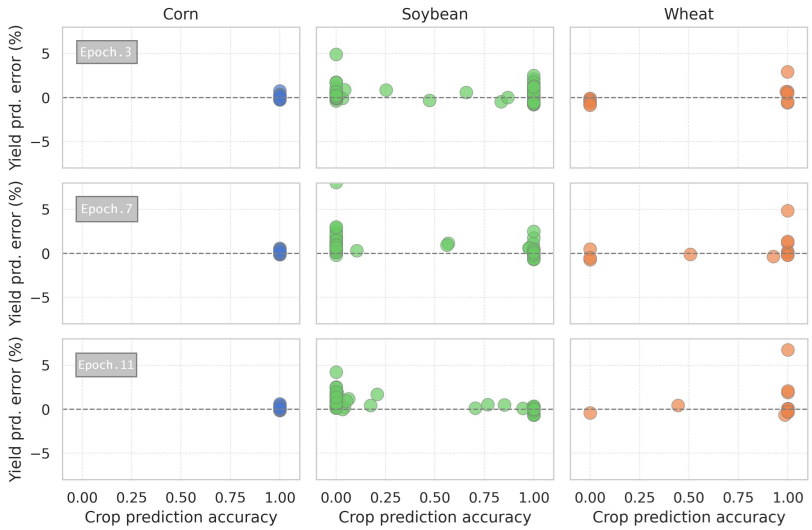


Fig. 4.2.: Comparison of model performance on the tasks of yield prediction and crop prediction for the CropYield dataset. The rows correspond to the results for epochs 3, 7, and 11, from top to bottom.

4.4.1 *CropYield*

Tasks performance per crop across fields

In the CropYield dataset, we evaluate the model performance in Experiment **10** across epochs, analyzing the relationship between yield relative error and crop prediction accuracy. Figure 4.2 presents results for epochs 3, 7, and 11, separately for each crop, with performance averaged per field. Each point in the figure represents a field, including training, validation, and test sets. We stop in this analysis at epoch 11 as the model achieved its best performance on the validation set at this epoch. The figure shows that corn fields consistently exhibit strong yield prediction performance and perfect crop classification accuracy, whereas the model faces greater challenges with the other two crop types. For soybean fields, a decrease in maximum yield prediction error is observed across epochs in fields with high crop classification accuracy, while fields with poor classification maintain high yield prediction errors, suggesting a correlation between correct crop classification and improved yield prediction. In wheat fields, fewer instances of poor crop classification are observed as training progresses.

Tasks performance per field across epochs

Since the results above are field-averaged, we further examine subfield-level performance by analyzing a random sample of pixels from two soybean fields in the test set. The results displayed in Figure 4.3 show the yield prediction relative error for correctly and incorrectly classified pixels throughout the training. In both fields, the yield prediction relative error is generally higher for misclassified pixels (orange) compared to correctly classified ones (blue), with this effect appearing in early epochs for one field and persisting after the model reaches optimal performance (epoch 11) in another. These findings suggest that incorrect crop classification at the subfield level negatively impacts yield prediction. Additionally, Figure 4.4 illustrates yield and crop type prediction

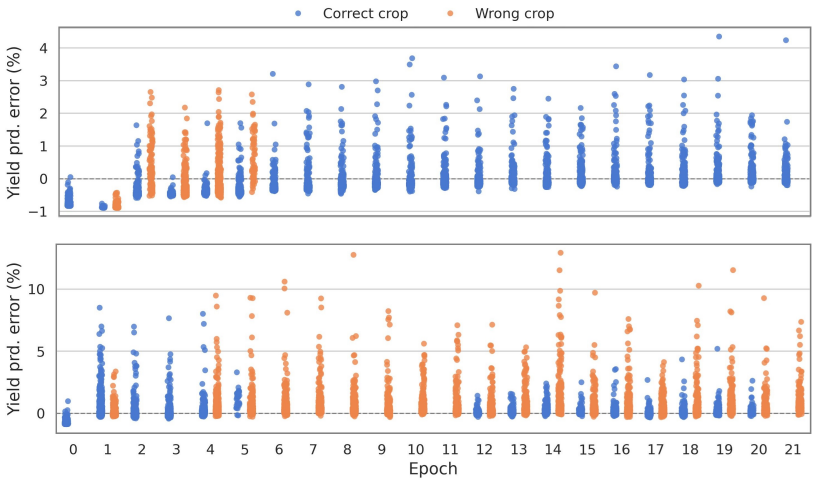


Fig. 4.3.: Comparison of model performance on the tasks of yield prediction (measured in relative error) and crop prediction accuracy for the CropYield dataset across 21 learning epochs. Results correspond to two soybean fields. 300 correctly classified and another 300 misclassified pixels are displayed for each field.

maps at different epochs, from fields where we clearly notice that regions with crop misclassification correspond to areas with significant yield underestimation.

4.4.2 *Benge*

Correlation analysis between main and auxiliary tasks

To investigate the explanatory potential of auxiliary tasks in the Benge dataset, we analyze Experiment **15**, which predicts all available modalities as auxiliary tasks (see Table 4.3). We first compute the Pearson correlation between the error of the main task, LULC classification, and the errors of the auxiliary tasks, on 10% of the test set. The results presented in Figure 4.5 indicate a decreasing correlation for all task combinations during early training epochs. While the LULC-Season correlation exhibits fluctuations throughout training, these variations are less pronounced in the LULC-Weather and LULC-ClimateZone combinations. In contrast, the LULC-DEM correlation remains more stable, likely due to the similar spatial resolution of both tasks, as they both return a single-channel image. This differs from the other auxiliary tasks, which predict tabular data. Although the correlations do not exceed 0.23, the p-values remain below 0.05.

Visual assessment of LULC & DEM correlation

To further examine this correlation between LULC and DEM, we present in Figures 4.6 and 4.7 five data samples where this relationship is clearly visible. Generally, we observe that errors tend to be correlated in regions where the model fails to accurately determine elevation, particularly along boundaries such as terrain edges or riverbanks. In these regions, land cover classification errors are more frequent. Conversely, when LULC misclassifications are scattered within a patch containing highly heterogeneous land cover, the correlation is weak. These areas typ-

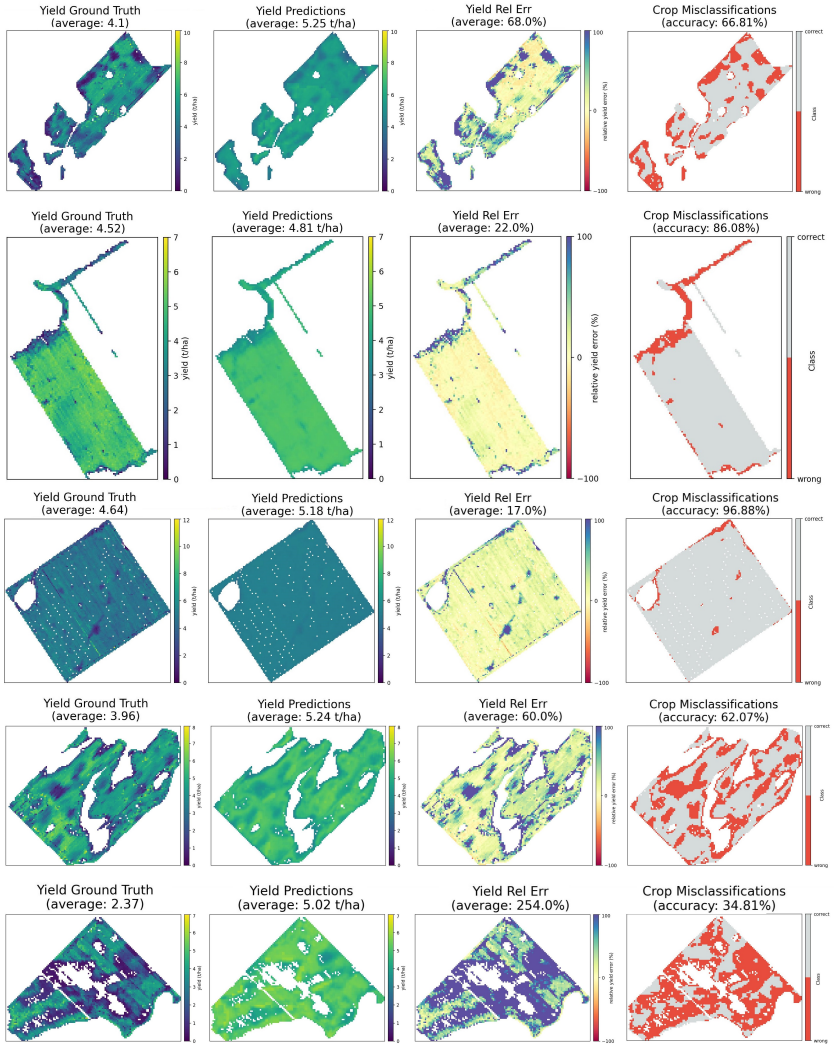


Fig. 4.4.: CropYield model performance on five soybean fields, at epoch 16 for the top field, epoch 4 for the following two fields, and epoch 14 for the remaining two. From left to right: Target yield, predicted yield, relative yield error, and crop misclassifications.

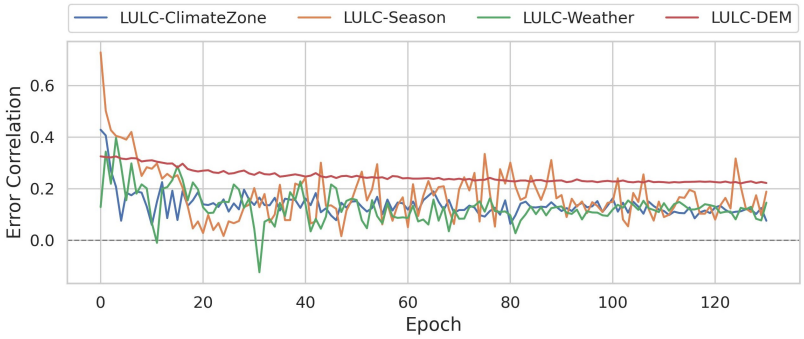


Fig. 4.5.: Error correlation between the main Benge task (i.e. LULC) and auxiliary tasks.

ically feature stable terrain elevation, leading to DEM prediction errors that do not exhibit the same scattered distribution as the land cover.

4.4.3 *TreeSAT*

Correlation analysis between L2 & L3 prediction accuracy

We investigate Experiment **7** in TreeSAT dataset, which predicts L2 and age alongside the main L3 label. We examine the correctness combinations of L2 and L3 labels in the test set throughout training, with results presented in Figure 4.8. Here, 'C' denotes a correctly predicted label, while 'F' indicates a false prediction. The notation follows the order of L2 and L3 predictions; for instance, 'CF' means that L2 was correctly predicted, but L3 was not. The results reveal that the count of instances where one label is correct while the other is incorrect (i.e., CF and FC) remain relatively stable throughout training. In contrast, the number of samples where both labels are correct (CC) consistently increases, while instances where both labels are misclassified (FF) decrease correspondingly. This trend suggests that FF samples are more likely to be corrected into CC as training progresses, whereas instances in which only one label is initially correct (CF or FC) are less likely to be fully corrected later during the learning process.

Analysis of the hierarchy in L2 & L3 predictions

Given the hierarchical nature of tree classes, we further examine how this structure influences the model's predictions. Figure 4.9 illustrates the distribution of L2-L3 prediction combinations and their adherence to the hierarchy at an early training epoch (Epoch 7) and at the best-performing epoch (Epoch 93). We add '-in' to the label in cases where the predicted L3 belongs to the predicted parent class L2, and '-out' to instances where it does not. The results indicate that when L3 is misclassified (i.e., in CF and FF cases), the proportion of instances where the predicted L3 remains within the predicted L2 class is consistently higher than those where it falls outside, regardless of whether L2 is correctly predicted. In other words, at both early training stages and the model's peak performance, CF-in is more frequent than CF-out, and FF-in is more frequent than FF-out. This suggests that the model has learned aspects of the hierarchical relationship between L2 and L3 and tends to respect it even when misclassifying L3. Note that in CC cases

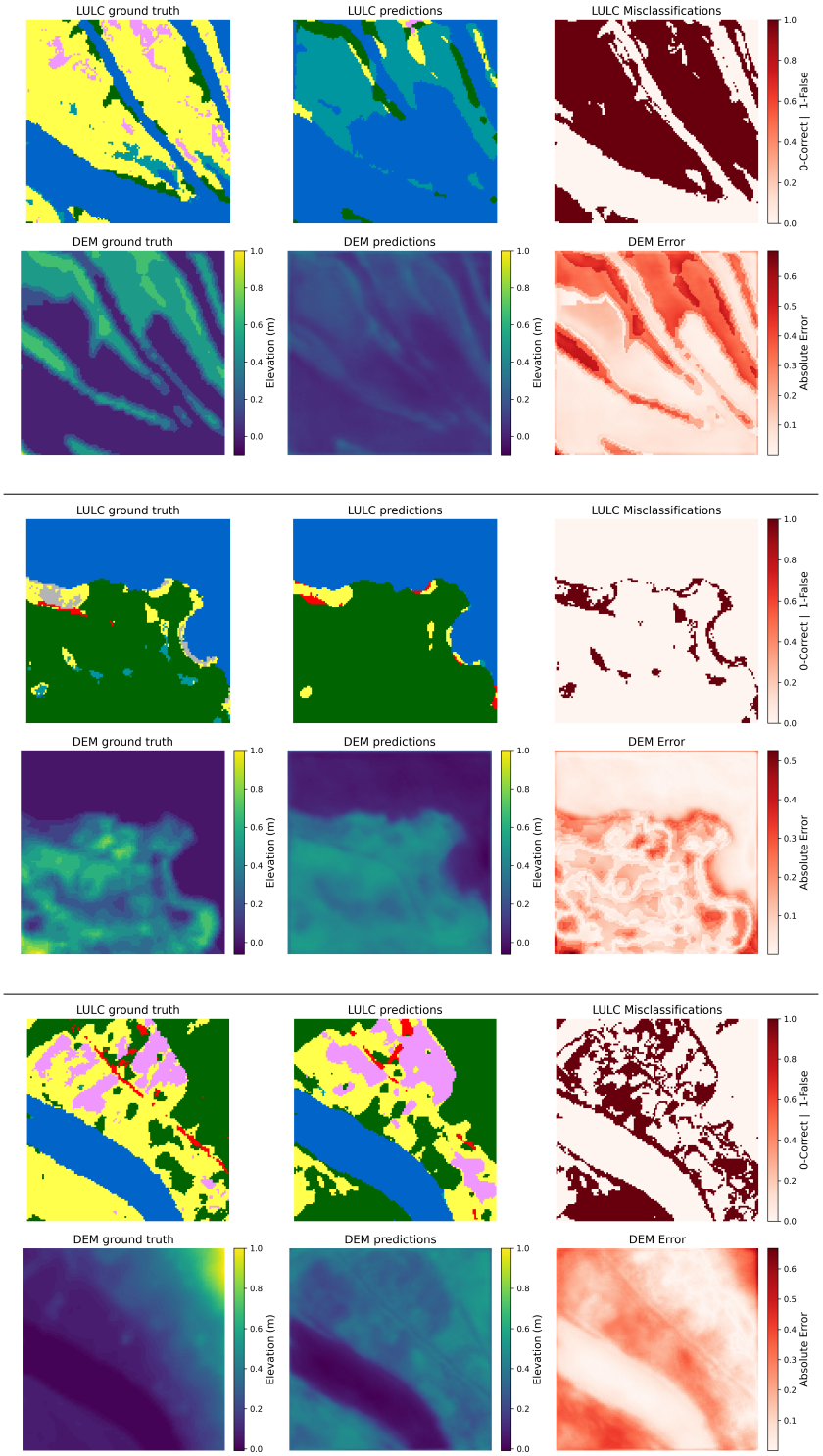


Fig. 4.6.: (1/2) Model predictions and errors, compared against the ground truths, on the LULC and DEM prediction tasks. The predictions are of the best epoch, on two random Benge dataset samples from the test set.

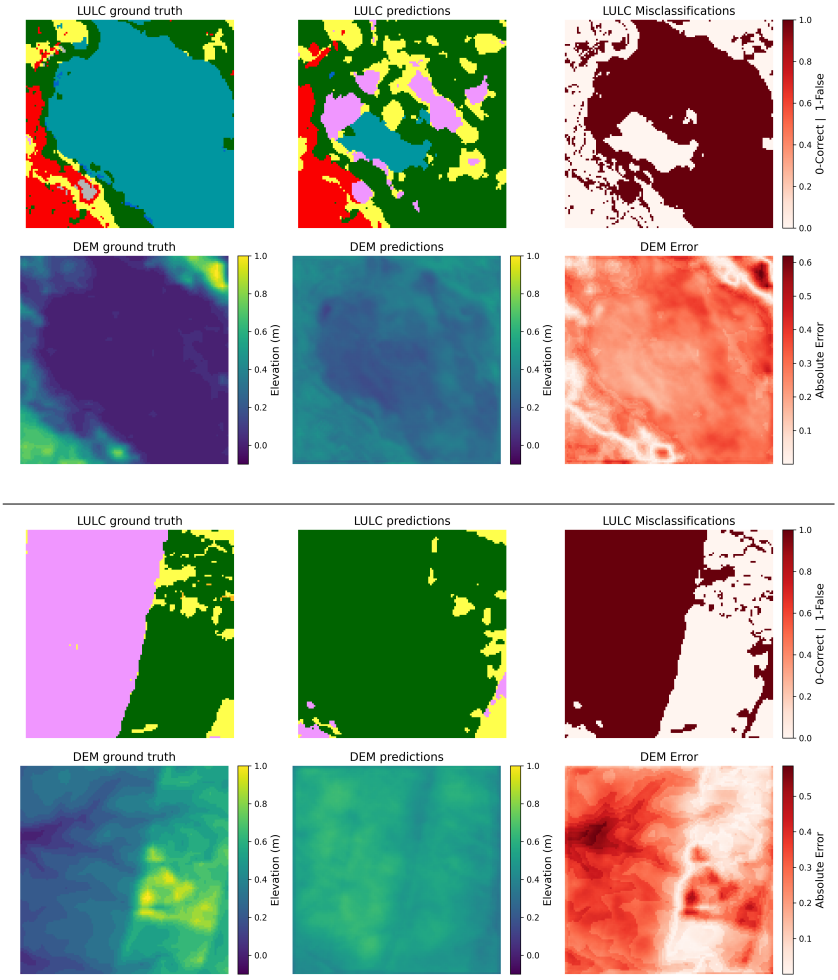


Fig. 4.7.: (2/2) Model predictions and errors, compared against the ground truths, on the LULC and DEM prediction tasks. The predictions are of the best epoch, on two random Benge dataset samples from the test set.

the hierarchy is always maintained, whereas in FC cases it is always violated.

Correlation analysis between age, L2 and L3 predictions

Since the experiment explained here also predicts the age, we further analyze the correlation between this modality and different L2-L3 correctness combinations. Figure 4.10 illustrates the average MAE for age prediction across different combinations of L2 and L3 prediction correctness. The results show that samples with both labels correctly predicted (CC) consistently achieve the lowest error scores throughout the training process. In contrast, samples with both labels incorrectly predicted (FF) consistently exhibit the highest error scores. The mixed groups, CF and FC, display fluctuating average MAE values, with CF showing lower error scores compared to FC. This suggests that an incorrect prediction of the L2 label has a negative impact on the accuracy of age prediction, more than an incorrect prediction of the L3 label.

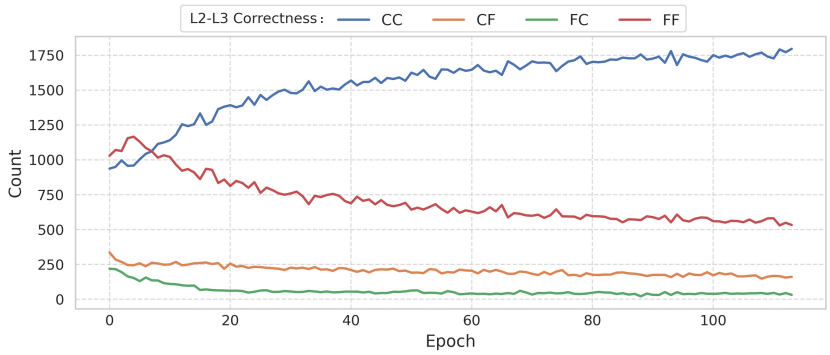


Fig. 4.8.: Count of combinations of correct or false classifications of L2 and L3 labels in the test set of TreeSAT dataset, throughout the training.

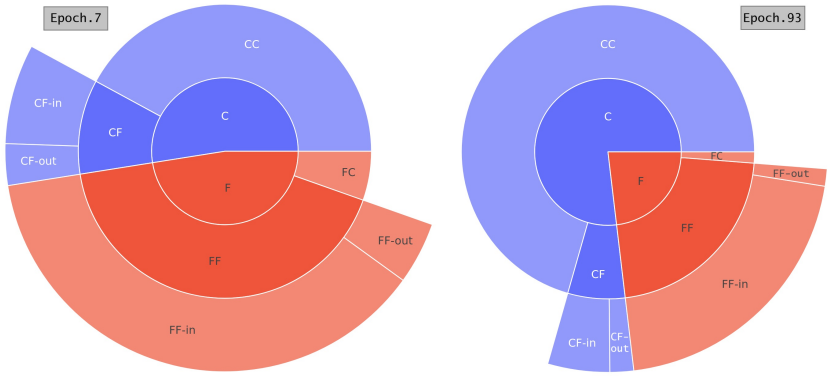


Fig. 4.9.: Pie Chart of the distribution of combinations of correct or false classifications of L2 and L3 labels. The results are inferred from the test set at epochs 7 and 93.

4.5 SUMMARY

In this chapter, we proposed a multitask learning framework to intrinsically enhance model interpretability. By leveraging multimodal datasets in RS, we exploited the rich information content of satellite data to shift additional input modalities into auxiliary tasks. This approach not only maintained comparable performance to baseline models but also reduced the need for the additional data sources at deployment. More importantly, it provided valuable interpretability insights, allowing us to analyze error correlations between the main and auxiliary tasks across three diverse RS datasets. We demonstrated how error propagation patterns could be identified and analyzed across different machine learning tasks.

Demonstrated benefits of the multitask setup

While our findings demonstrate the potential of multitask learning for model interpretability, there are several promising directions for follow-up work. The correlation patterns identified through the analysis of error maps in Benge and CropYield were observed in a limited number of samples. However, the presented examples provide evidence of error propagation between the main and auxiliary tasks, suggesting that follow-up work could leverage this approach to enhance performance in both tasks. For instance, integrating insights from the interpretability

Follow-up work

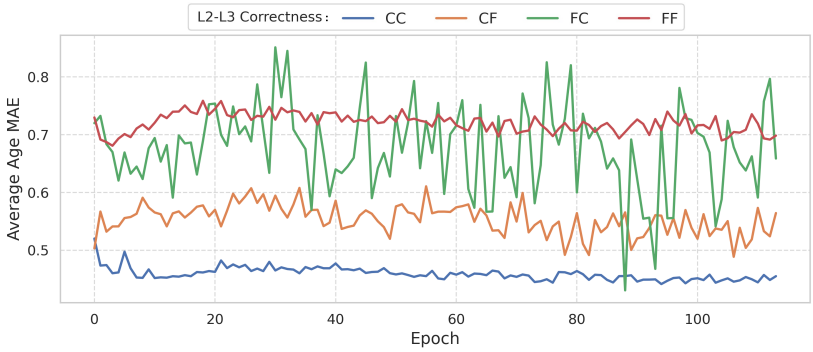


Fig. 4.10.: MAE of the age prediction task, averaged across each combination of correct or false classifications of L2 and L3 labels in the test set of TreeSAT dataset, throughout the training.

Constraining the loss function

Neural architecture search

Task weighting in multitask learning

analysis as constraints within the loss function could enforce meaningful relationships between tasks. Using the hierarchical structure of labels in the TreeSAT dataset to refine predictions is one example. Moreover, optimizing the neural architecture could further improve our results, aligning with findings from prior studies in which multitask learning outperformed single-task baselines [203, 63, 198, 160, 268, 190, 179]. Finally, another promising direction is to refine the selection of task weights in multitask learning. Automating this process using uncertainty estimation [160] or adaptive weighting based on loss improvement rates [190] could enhance the balance between tasks.

PartIII

XAI FOR IMPROVEMENT

5 Sufficient and Necessary Features

5.1 INTRODUCTION

5.1.1 Data-Centric vs. Model-Centric Machine Learning

The Machine Learning cycle

Recent efforts within the EO research community have primarily focused on enhancing model architectures and training strategies to boost performance, and significant advancements have been made. However, ML is a cyclical process that extends beyond model design and training. As depicted in Figure 5.1, the ML cycle involves the problem definition, data collection and preparation, model training and evaluation, and final deployment, which informs an improved or a new problem definition. While the model-centric approach emphasizes the training and evaluation stages, the data-centric approach focuses on the remaining three steps.

From model-centric to data-centric ML

Therefore, as a complementary component to these model-centric efforts, research is currently increasingly diverging towards a more data-centric approach [202, 256]. This shift aims to better address challenges faced during data acquisition and data curation stages. Subsequently, the enhancement of the quality of the input data can improve the performance and reliability of the models. Furthermore, incorporating a feedback loop that includes model evaluation results can provide valuable insights for refining both the data and the models.

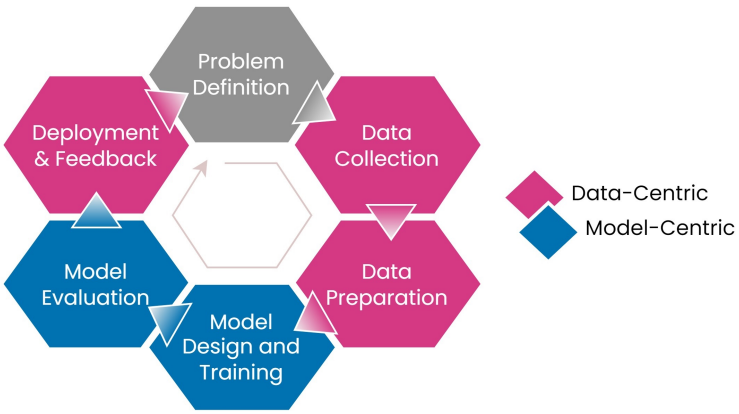


Fig. 5.1.: ML cycle. Reproduced and modified from [256].

Tab. 5.1.: Summary of traditional feature selection methods: their mechanism and main limitation.

	Filter methods	Wrapper methods	Embedded methods
Mechanism	use statistical properties to infer relevance to the target variable.	evaluate feature subsets based on the performance of a specific ML algorithm (e.g, SVM).	perform feature selection during the model training process, often through regularization techniques (e.g, LASSO).
Cons	independent of the model training.	extracts feature relevance from a model different from the main one.	the shrinkage leads to biased estimates of the coefficients.

5.1.2 *Feature Selection Techniques*

Data-centric ML via feature selection

Within the scope of data-centric mML, this chapter considers feature engineering techniques, particularly feature selection methods. These techniques aim at identifying the most useful and necessary features relevant to the task at hand. The objective is to avoid supplying the model with redundant information or extraneous features of the available input data that may cause the model to learn spurious correlations and hinder its capacity to generalize. The prevailing belief that "*more data is better*" does not always hold true, and often comes at the expense of immense computational resources usage and a heavy impact on the environment [79].

Traditional feature selection methods

Traditional feature selection methods, such as filter, wrapper, and embedded approaches, often involve exhaustive search strategies that can be computationally prohibitive for large datasets. **Filter methods** are pre-processing steps independent of the model training. They select features based on their statistical properties and relevance to the target variable. **Wrapper methods** evaluate feature subsets based on the performance of a specific ML algorithm. One prominent work in this domain is the Recursive Feature Elimination (RFE) algorithm, which recursively removes the least significant features based on their importance weights, as determined by a SVM [118]. In EO, Zhang et al. [345] use a similar technique based on feature importance extracted from a random forest model to identify the most relevant features in marine data. A major limitation of this approach is that the feature importance is extracted from a model different from the main one, while these scores are usually model dependent. **Embedded methods** perform feature selection during the model training process, often through regularization techniques. Examples include the Least Absolute Shrinkage and Selection Operator (LASSO) method [302], which performs feature selection by enforcing sparsity through L1 regularization, and implicitly removes less important features during model training. Table 5.1 summarizes the three types of methods described above and their main limitation.

Incremental deletion of feature in XAI

In the field of XAI, the incremental deletion approach is also used to evaluate the correctness of feature importance scores, i.e. how faithful these scores are to the model [232]. Given that deleting features by setting them to zero, for instance, can lead to out-of-distribution samples, an alternative is to retrain the model on the modified data. Specifically, Hooker et al. [136] propose the RemOve And Retrain (ROAR) method,

which evaluates the feature attribution estimates by removing the $k\%$ most important features, and then retraining the model.

5.2 EXPLAINABLE AI FOR FEATURE SELECTION

5.2.1 Methodology

In this chapter, we use the ROAR method to identify the set of sufficient and necessary predictive features to achieve a "*good*" model performance. We focus on temporal multi-source RS datasets. In particular, we leverage feature attribution methods to estimate how much each feature contributes to the final predictions. Subsequently, by employing an incremental deletion approach, we iteratively remove less important features until an optimal set of predictive features is identified. This method ensures an active optimization of the data at each cycle, leading to a final model that efficiently exploits the available data.

Incremental deletion based on feature attributions

Unlike the original implementation of ROAR in [136], which replaces the features-to-delete with a constant value (i.e., masking features instead of deleting them), in our study we completely remove the feature before adjusting and retraining the model. This setup allows the simulation of a scenario where the deleted features are genuinely absent from the initial dataset. Such an approach is particularly valuable when it reveals the unnecessity of certain modalities to the task, thereby reducing the future workload associated with data collection, cleaning, and pre-processing for practitioners. Moreover, analyzing the deletion results across time steps helps in assessing the potential impact of missing data instances, such as those caused by cloud cover or satellite failure. Furthermore, our deletion approach can potentially provide valuable insights into the structure of the data, the potential redundancy among features, and their overall contribution and necessity to the ML task at-hand.

Deleting instead of masking features

It is important to note that the "necessity" and "sufficiency" of a feature are properties defined relatively to the explained model. A feature is considered necessary for the model if it receives an attribution score higher than other features and its deletion results in a significant performance drop after retraining. Conversely, a sufficient feature subset is one on which the model, when trained exclusively, can achieve performance comparable to the baseline model trained on the full feature set.

Defining necessary and sufficient features

Our framework can be summarized as follows: First, we select a highly performing model by comparing recurrent, convolutional, and attention-based networks. Subsequently, we incrementally delete either the most or least important features based on importance scores estimated by six different methods and retrain the selected model. We conduct separate feature deletion operations on the spectral and temporal dimensions to identify the necessary modalities and time steps, respectively, and validate our approach on multiple temporal and multimodal geospatial datasets. Figure 5.2 illustrates the overall workflow of our approach.

Overall workflow of the proposed framework

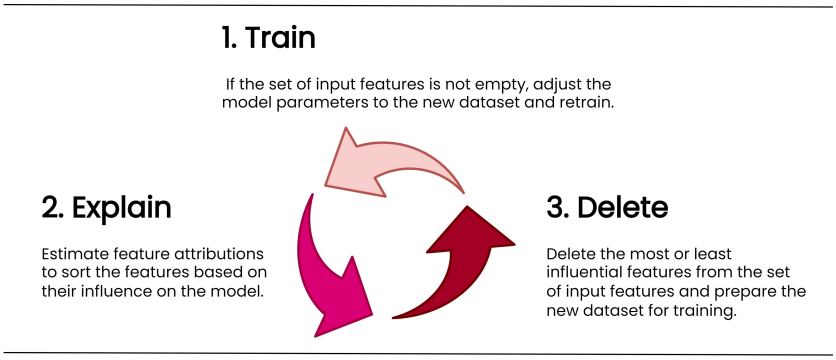


Fig. 5.2.: Overall flow of the incremental deletion approach based on feature attribution estimations.

5.2.2 *Datasets*

We apply the incremental deletion framework to the following three EO datasets, which are composed of different modalities and span multiple years, covering both regression and classification tasks.

CropHarvest. The CropHarvest dataset is a multimodal temporal dataset designed for crop classification tasks [308]. The input modalities include satellite data from S2 and S1, weather time series, and static topographic information. The temporal modalities are provided on a monthly basis over multiple years (2016 - 2022). We use a multi-crop version of this dataset, which includes 10 classes.

SwissYield. For crop yield estimation, we use a dataset specifically aimed at predicting cereal crop yield [246]. This regression task involves predicting the amount of crop (in t/ha) grown in a large farm in western Switzerland during the growing seasons from 2017 until 2021. The input modalities available in this dataset include multispectral satellite data from S2 mission and weather time series. The weather data is aligned with the satellite temporal resolution of five days, covering the period from seeding to harvesting.

China PM2.5. The China PM2.5 dataset provides data for the prediction Particulate Matter (PM) concentrations, specifically PM2.5, across various regions in China, to track and monitor air pollution [46]. This dataset covers the years from 2010 until 2015 and includes multiple modalities, namely weather information and ground-based air quality measurements. The task is to predict PM concentrations using hourly measurements from the preceding 7 days leading up to the date of the target value.

5.2.3 *Modeling*

Data splitting strategy

Each dataset is split into three subsets: the training set comprises data from all years except the last two, which are uniformly divided into validation and test sets. This strategic split aims at reproducing the

Tab. 5.2.: Best model performance for each architecture. The best score in each dataset is highlighted.

Model	CropHarvest (Accuracy)		SwissYield (R^2)		China PM2.5 (R^2)	
	Validation	Test	Validation	Test	Validation	Test
MLP	62.1	61.9	52.4	42.1	68.0	54.3
RNN	65.7	65.5	44.0	33.5	47.3	44.1
LSTM	56.8	56.9	43.3	35.6	63.0	57.1
GRU	63.6	63.6	54.6	47.8	71.5	64.1
TempCNN	**87.8**	**86.6**	**67.4**	**61.8**	**77.5**	**67.2**
TAE	67.9	67.9	58.5	51.0	66.3	57.5
LTAE	67.3	67.2	63.1	57.2	72.4	62.4

common case where real-life application can only train the models on data from previous years to be deployed for upcoming seasons.

Evaluation of multiple model architectures

To identify the most effective architecture for processing multivariate time series data, we conduct a comparative evaluation of seven different model architectures. Each architecture is tested under multiple hyperparameter settings and evaluated on the validation set to determine its performance. The models compared in this study include a MLP, a basic RNN and two variants: LSTM [131] and gated recurrent Unit (GRU) [50], a temporal convolutional neural networks (TempCNN), and finally two attention-based models: temporal attention encoder (TAE) [98] and lightweight-TAE (L-TAE) [96].

We trained each model architecture with various configurations and compared their performance on the validation set. Table 5.2 reports the results of the best-performing configuration in each model architecture. The metrics reported are the accuracy for the classification task and R^2 score for regression tasks, evaluated on the validation and test sets. The results indicate that TempCNN consistently achieves the highest scores across all datasets, demonstrating its efficiency in handling both short and long time-series data. Attention-based models achieve the second-best scores, followed closely by the GRU network. Interestingly, the MLP outperformed the LSTM network in all datasets, as well as the RNN in the two regression tasks.

Results discussion

5.2.4 *Attribution Estimators*

Feature attribution methods are XAI tools that aim to quantify the contribution of input features to the output of a ML model. These methods provide insights into which features are influential in making predictions. In this chapter, we compare perturbation-based and gradient-based methods to infer feature attributions, namely SVS [286] and Guided Backprop (GB) [282] methods. SVS was presented in previous chapter, particularly in Section 3.2.1. GB overrides gradients of ReLU functions so that only non-negative gradients are backpropagated.

Base attribution estimators

Additionally, two ensemble-based variants of each base estimator are implemented: SmoothGrad-Squared (SG-SQ) [277] and VarGrad (VAR) [4]. SG-SQ averages a set of N noisy estimates of feature importance (i.e. Gaussian noise is injected in the input independently N times, before estimating the feature importance). The estimates are squared before being averaged. VAR uses a set of N noisy estimates of feature importance

Ensemble-based attribution estimators

as well, but computes their variance instead of the average. As shown in [136], these ensembling methods can strongly improve the correctness of the attributions of the base estimators.

Selection of explanation samples

To ensure that the analysis captures the overall influence of each feature, all attributions are considered in their absolute values. A random selection of 5000 samples from the training set is used to estimate the attributions, which are averaged to derive the ranking of the features. The same selection of samples is used across all experiments of the same dataset.

Feature grouping in SVS

Moreover, we use a feature grouping strategy applicable only on the SVS method to estimate temporal and band importance, by considering collective contributions to the model's predictions. Temporal importance is estimated by grouping the bands at each time step, perturbing them together to infer the significance of that particular time step. Similarly, band importance is assessed by treating the time series of each band as a group.

5.2.5 *Incremental Deletion Process*

Baseline Model After evaluating multiple architectures as described in 5.2.3, the model with the best metric score on the validation set is selected as the baseline model for the incremental deletion cycles. The metric score is defined based on the optimization task: accuracy for classification and R^2 for regression. The validation loss is used for early stopping.

Deletion Order We conduct two types of feature deletions: either based on the most important features or the least important. Progressively deleting the most important features can reveal the necessity of the key features to reach the baseline performance. Conversely, eliminating features that do not significantly contribute to the predictions can reduce noise in the input data, potentially enhancing model performance. This method can also identify a subset of features that are sufficient to reach the baseline performance after all extraneous features have been eliminated. The number of features deleted in each step depends on the dataset. By default, the deletion process addresses a single feature at a time. For long time-series, a larger step is used.

Deleting most important features first

Deleting least important features first

Cycles After training the baseline model, feature attributions are estimated using six different estimators (two base estimators + two ensemble-based variants of each), grouped by time-steps or bands, and the corresponding features are ranked accordingly. A new copy of the dataset is created after deleting the most or least important features, and a new model instance is trained with this modified data. The model architecture remains consistent, except for the modifications necessary to handle the new input size. For instance, removing a spectral band from the satellite modality would require adjusting the number of input channels in the first convolutional layer. Post-training, the new model is explained, and the attribution scores averaged over the selected samples are used to rank the features and decide which ones to delete in the next cycle.

Feature attributions estimation & ranking

Feature deletion, dataset & model adjustment

Repeat the cycle

This process is repeated, updating the feature attribution estimates and modifying the input data after each training, until only a single (or a set of) feature(s) is left in the addressed dimension at the final cycle.

5.2.6 *Incremental Deletion Results*

Base Attribution Estimators

We apply the incremental deletion approach on each dataset using the TempCNN architecture. We evaluate the model after each cycle and report its performance results for band and time-step deletion on the validation set, as shown in Figures 5.3 and 5.4, respectively, using the base attribution estimators. A horizontal line indicates the baseline performance in each plot.

Deleting the most important bands first

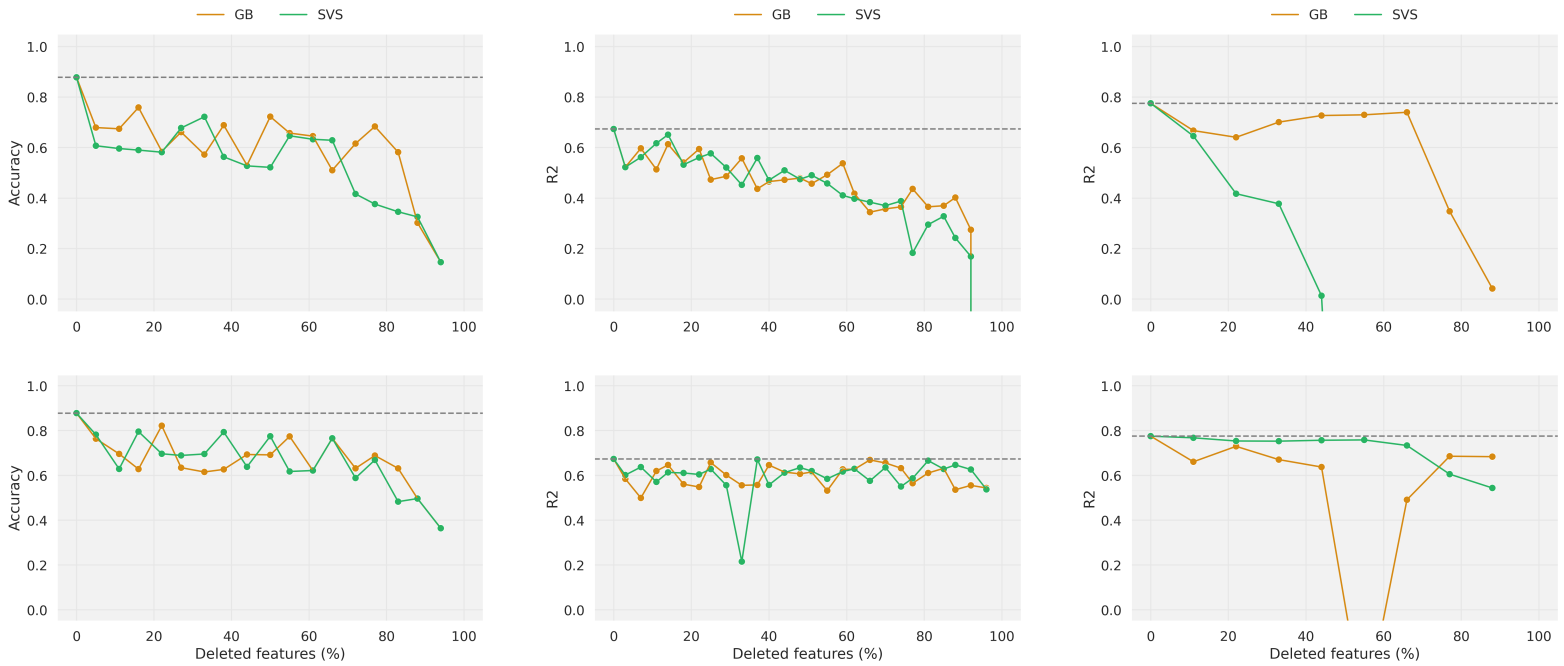
In the top row of Figure 5.3, the most important bands are deleted first. We observe that the performance progressively declines in the CropHarvest and SwissYield datasets. The decrease in performance is more significant for the PM2.5 dataset, especially when using the SVS attribution estimators to rank the features. This observation has two implications: first, the correctness of the attributions returned by SVS exceeds those provided by the GB method; and second, the baseline performance in the PM2.5 dataset relies on the two most important features (wind speed and direction), with the remaining features being insufficient for the model to achieve the baseline performance. In contrast, in the agricultural datasets, many important features can be dropped before a significant decline in model performance is observed. This also indicates that these features are not necessary for achieving a performance comparable to the baseline.

Deleting the least important bands first

Deleting the least important bands, as shown in the second row of Figure 5.3, reveals additional insights. In the CropHarvest dataset, removing up to 70% of the least important bands does not reduce the model accuracy below 60%. In the SwissYield dataset, the model can still recover its baseline performance after more than 80% of the bands have been removed according to the SVS method. Notably, the final instance in the corresponding plot demonstrates that training on the time series of a single band still yields high performance. The specific band varies depending on the attribution method used: SVS retains B11, a SWIR satellite band, whereas GB selects B06, a RE band of the same satellite. Using the same band deletion approach, the model trained on the PM2.5 dataset maintains its baseline performance even when 65% of the features are deleted. In this case, wind speed, wind direction, and humidity conditions were sufficient for achieving the baseline optimal performance.

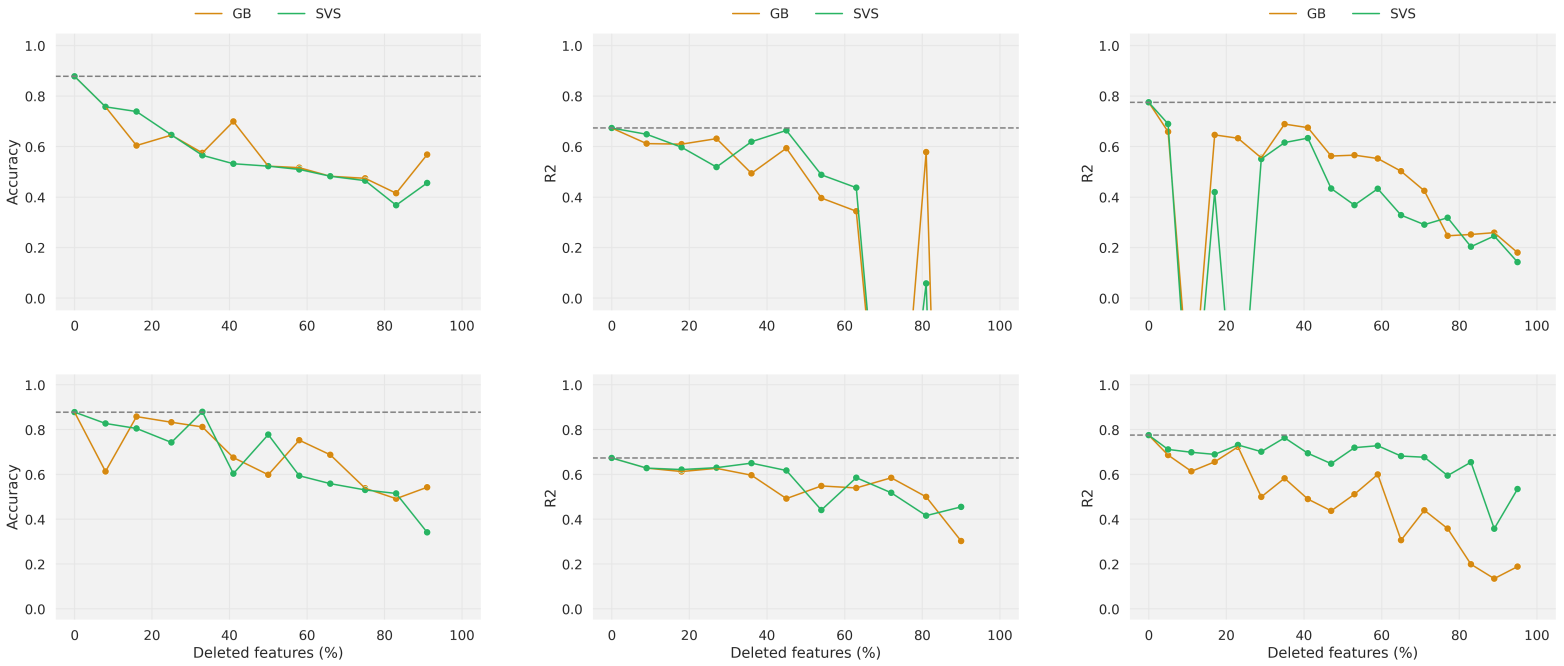
Deleting the most important time-steps first

The results of the time-step deletion analysis provide insights into the time periods whose absence significantly impacts model performance and those which are sufficient to achieve baseline accuracy. In Figure 5.4, removing the most important instances first, as shown in the top row, results in a consistent decline in performance in the CropHarvest dataset, particularly when using the SVS estimator. For the PM2.5 dataset, the R^2 scores with the SVS estimator are lower than those achieved with the GB estimator, especially when more than 30% of the time-steps are



(a) CropHarvest (b) SwissYield (c) China PM2.5

Fig. 5.3.: Incremental band deletion results, comparing SVS and GB attribution estimators. In the top row the most important bands are removed first, and the least important bands in the second row.



(a) CropHarvest **(b)** SwissYield **(c)** China PM2.5

Fig. 5.4.: Incremental time-step deletion results, comparing SVS and GB attribution estimators. In the top row the most important time-steps are removed first, and the least important time-steps in the second row.

removed. The behavior of the two estimators is mixed in the SwissYield dataset. In all datasets, the moderate slope of the curves indicates that the information required by the model for accurate prediction is distributed across multiple instances, rather than being concentrated in a few critical time-steps.

Deleting the least important time-steps first

The results of removing the least important instances first in the second row in Figure 5.4 show that the model can still perform similarly to the baseline after deleting more than 30% and 40% of the time-steps in CropHarvest and SwissYield, respectively. In PM2.5, a performance comparable to the baseline can be achieved even when more than 80% of the instances are deleted, according to the feature ranks provided by SVS. Taking a closer look at the time-steps left at this point revealed that the hourly instances from the last two days (within the 7-days window) were sufficient to reach a high R^2 score.

Ensemble-based Attribution Estimators

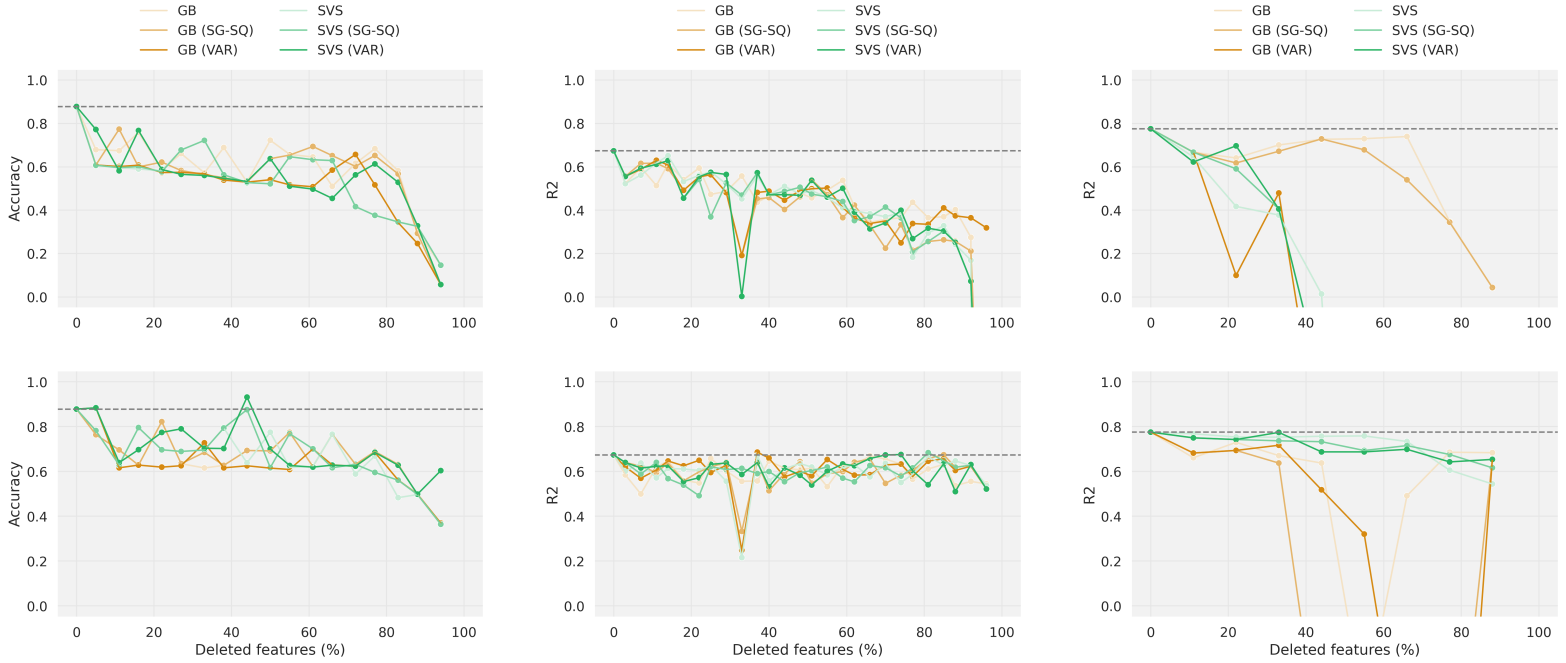
We repeat the experiments described above using the ensemble-based variants. The results of the incremental deletion based on SG-SQ and VAR methods applied to each estimator are illustrated in Figures 5.5 and 5.6. Hooker et al. [136] demonstrated how these variants significantly improve the correctness of gradient-based attribution estimator. In our experiments, the only noticeable improvement is observed in the PM2.5 dataset when removing the most important bands first. Specifically, GB (VAR) shows a significant decline in model performance, reflecting a correct estimation of the feature importance ranking. Another noteworthy improvement is observed in the SwissYield dataset, where removing the least important time-steps first, as displayed in Figure 5.6.(b), shows a high accuracy even in the last cycle of using the SVS (SG-SQ) method. This suggests a filtering of the least important features that is faithful to the model and its reasoning. Concretely, in this dataset, the first SWIR band from S2 satellite data was sufficient for achieving baseline performance.

Cases of improved attribution estimations

5.3 SUMMARY

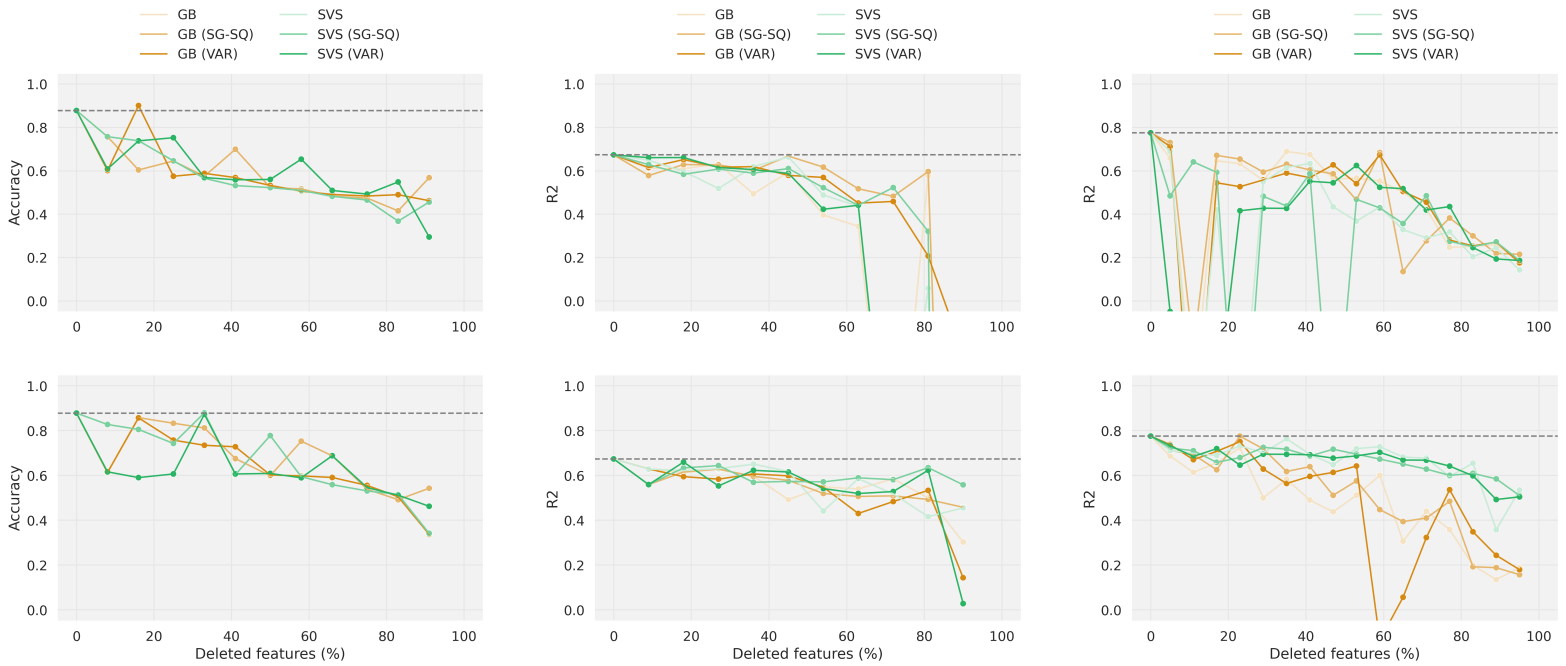
Inspired by the ROAR framework, we proposed an efficient approach to identify a small subset of bands and time-steps in geospatial temporal data sufficient to reach the model's baseline performance, i.e. the accuracy reached when providing the model with all available modalities and instances. We evaluated this approach on three datasets, and showed how many features can efficiently be removed before a significant drop in accuracy is observed. Additionally, we found that in some datasets, performance declines immediately after a few features identified as the most important are deleted. This suggests that these features are necessary for the baseline performance and that the information they encode is absent in the remaining features.

Furthermore, the expected behavior regarding the decline in performance when starting with the deletion of the most or least important features also revealed a higher correctness of the attributions estimated



(a) CropHarvest **(b)** SwissYield **(c)** China PM2.5

Fig. 5.5.: Incremental band deletion results, comparing the ensemble-based estimators SG-SQ and VAR, applied to SVS and GB. In the top row the most important bands are removed first, and the least important bands in the second row.



(a) CropHarvest **(b)** SwissYield **(c)** China PM2.5

Fig. 5.6.: Incremental time-step deletion results, comparing the ensemble-based estimators SG-SQ and VAR, applied to SVS and GB. In the top row the most important time-steps are removed first, and the least important time-steps in the second row.

by SVS compared to GB. The ensemble-based variants only improved the faithfulness of GB estimates in a few cases.

This work can be enhanced by comparing additional feature attribution methods. The faithfulness of the chosen method enhances its ability to identify a minimal feature subset necessary to achieve the model's baseline performance. Moreover, comparing these results across various model architectures might help identify the features necessary and sufficient for predicting the target regardless of the model employed.

6 Vegetation Indices

6.1 INTRODUCTION

6.1.1 Vegetation Indices in Earth Observation

A common data engineering practice in EO is the usage of VIs. VIs allow an efficient monitoring of vegetation growth and agricultural activities. Ratios, differences, and derivatives between reflectance values from different spectral wavelengths can enhance the spectral signals associated with vegetation characteristics of interest, given that the original measurements of spectral reflectance constitute a mixed signal comprising vegetation canopies, shadows, soils, and other components present on the land surface [341]. While some VIs, such as the NDVI [257], the Enhanced Vegetation Index (EVI) [139], and the Soil-Adjusted Vegetation Index (SAVI) [140], are commonly used for crop monitoring, the selection of the most suitable index is not always straightforward [341]. Instead, the initial step involves identifying the sensitive wavelengths and corresponding VIs for their optimal use.

Common indices

Previous generations of satellites were capturing a limited number of spectral bands, and a few expert-designed VIs were sufficient to harness their potential. New generations of multi- and hyperspectral satellites can however capture additional bands, but are not yet efficiently included in indices. For instance, the S2 multispectral instruments stand out as one of the few remote sensors with the capacity to capture RE wavelengths between 700 and 800nm. Notably, the additional RE bands remain underexplored for their potential to enhance crop classification through VIs [218]. Furthermore, SWIR bands, typically used for water monitoring, have also received little attention in exploring their efficacy to track vegetation cover and its phenology [218].

Unexplored bands in new generation satellites

6.1.2 Vegetation Indices in Deep Learning

An advantage of using DL lies in the model's inherent capability to automatically extract crop-related features and discern interactions between raw bands [81]. Unlike classical ML models which usually rely more heavily on VIs, a DL model is assumed to not require explicit feature engineering, due to its capacity to learn complex patterns and representations directly from the raw input data. We have also demonstrated in Chapter 3, Section 3.2.4, how the model trained on the S2 bands outperformed the model trained on the selected VIs.

Automated feature extraction in DL

In contrast, as also mentioned in Section 3.2.4, VIs offer an advantage over raw satellite bands due to their initial design for agricultural applications, making them more interpretable. Additionally, we argued in Section 3.2.5 that the suboptimal performance of models trained on

Potential of using VIs

indices might stem from inefficiently selecting these indices from the extensive list available in the literature.

Our contribution

In this chapter, we challenge prevailing assumptions and compare the performance of a DL model trained directly on raw satellite bands versus VIs selected based on XAI guidance. Specifically, we first train a deep neural network using multispectral satellite data, then estimate feature attributions to identify the most influential bands. We subsequently either choose existing VIs incorporating these bands or create new ones before retraining the model using only index-based inputs. We validate our approach on a crop classification task.

6.2 SELECTION AND DESIGN OF VEGETATION INDICES

6.2.1 Dataset & Experimental Setup

Crop Mapping Dataset

Crop classification for Sub-Saharan regions

In Sub-Saharan Africa, extreme food insecurity and malnutrition are prevalent in multiple countries. Addressing these challenges involves implementing more efficient agricultural systems and undertaking regional monitoring of harvested crops. VIs offer a valuable tool for distinguishing between crops and estimating their health and growth stages [164].

EO dataset for crop classification in Ghana and Sudan

In this study, we leverage multispectral satellite data of Ghana and South Sudan to address this task. The corresponding public datasets used contains S2 satellite image time series captured between January and December 2016 at a 10m resolution, and are labeled with multiple land cover classes [260]. For our study, we merge datasets from both regions and retain only the pixels corresponding to crops. We focus our work on classes with more than 10,000 labeled pixels, which include sorghum, maize, rice, groundnut, soybean, and yam. Table 6.1 presents a summary of the data distribution in each country. We partition 5% of the data for validation, ensuring that pixels originating from the same satellite image patch are exclusively used for either training or validation but not both.

Tab. 6.1.: Pixel count per crop type in Ghana and South Sudan (S-Sudan) datasets.

Crop	Ghana	S-Sudan	Total
Maize	322,767	7,080	329,847
Groundnut	96,371	4,943	101,314
Rice	93,908	5,078	98,986
Soybean	67,638	0	67,638
Sorghum	8,352	56,833	65,185
Yam	22,091	0	22,091

Modeling

Input features

We use ten bands from S2 data for our analysis, namely the blue (B02), green (B03), red (B04), three RE bands (B05, B06, B07), NIR (B08), narrow Near-InfraRed (n-NIR) (B8A), and two SWIR (B11, B12) bands.

An additional channel, indicating the cloud coverage of the image, is stacked to these bands and used in all our experiments.

DL model and training setup

Regarding the modeling, we rely on recurrent networks, which have successfully been used to analyze temporal satellite data [150, 272, 95, 225]. We opt for the GRU, introduced in [50], due to its moderate number of parameters and its proven effectiveness in remote sensing applications [95, 144, 226]. The time series of each pixel are pre-padded to a fixed sequence length of 228, to account for the longest time series in the dataset, before being supplied to the model pixel-wise. To handle the unbalanced labels in the data, we use a weighted sampler during training. It assigns higher probabilities to small classes over large classes, enabling the model to train on a similar number of samples from each class during each training epoch. We evaluate all models using the overall accuracy (OA) and macro-F1 scores. The OA is the percentage of correctly classified pixels across all classes. We also report the accuracies per class.

Exploiting Spectral Attributions

Estimation of spectral attributions using SVS

Feature attribution methods are explanation techniques which assign a contribution score to each input feature, quantifying its relative importance to the model's prediction [200]. In this chapter, we use the SVS to estimate feature attributions [286]. SVS is grounded in cooperative game theory, which provides a solid theoretical foundation, unlike many other methods [200]. Its robustness has being evaluated in chapters 3 and 5 in the context of regression and classification tasks based on time series of satellite data, and has shown superior stability against several other techniques.

Overall flow of the proposed framework

The results of the spectral attribution (i.e., attribution scores of the satellite bands) are used to improve the selection of VIs for the crop mapping task in the following steps: We first interpret the model trained on the ten satellite bands by estimating the attributions for a maximum of 10,000 correctly classified pixels from each crop. The features are grouped over the spectral dimension to compute a single attribution value for the time series of each band. The negative attributions are suppressed to only consider positive contributions to a given class [267]. To standardize the results, we scale the attributions so that the summation of attributions per pixel equals 1, before averaging them both per class and globally. Subsequently, we use these importance values to select VIs that account for these bands, and adjust existing indices as needed. The model is then retrained by replacing the satellite bands with individual indices or binary combinations.

6.2.2 *Spectral Attribution Results*

Performance of the baseline model

We train the GRU-based model using the satellite bands and present the evaluation results on the validation set in the second column of Table 6.2.3. This baseline model achieved a score of 67% on both the OA and F1 metrics. In individual classes, high accuracies of 84% and 86% were attained for rice and sorghum, respectively, while yam exhibited the lowest score at 27%. This could be attributed to the relatively small number of pixels in this class. Notably, the largest two classes did not

necessarily exhibit the best performance, suggesting that the performance gaps are not solely due to the size of each class, given that we used a weighted sampling strategy to mitigate the class unbalance issue.

Global attribution results

We interpret the baseline model following the procedure described in the previous section, and visualize the corresponding results in Figure 6.1. Starting with the global average attribution line in black, SWIR-1 and RE-1 rank at the top with around 20% of the total importance, followed by the red, SWIR-2, n-NIR, and NIR bands, in the descendant order of their respective importance. The remaining bands exhibit a less significant importance. Notably, the relatively small importance of the cloud mask across all classes indicates that the model is not biased by this channel for the identification of any specific crop.

Crop-specific attribution results

Examining the crop-specific attribution results, we observe that groundnut and soybean highly rely on the first RE band, followed by the red and SWIR-1 bands. Sorghum has a similar attribution pattern. Rice has an additional particular dependence on the SWIR-2 band. Rice and yam identification significantly relies on the first SWIR band, followed by RE-1. All the remaining bands have each less than 10% of the total importance. Maize crop classification is sensitive to the first SWIR and RE bands, followed by the red band.

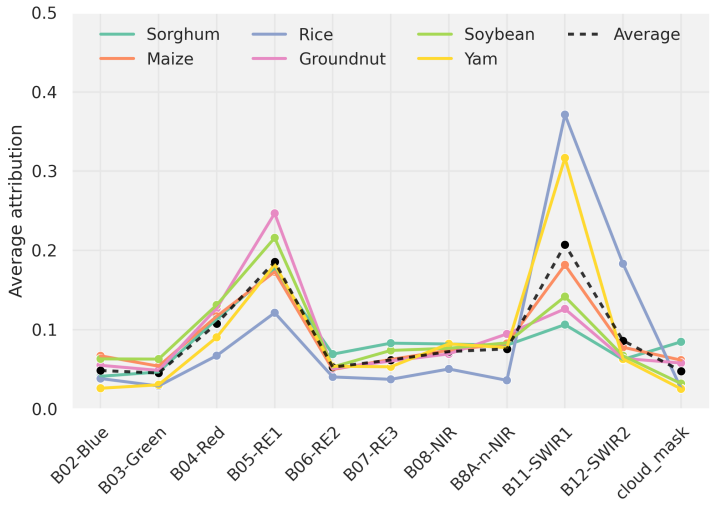


Fig. 6.1.: Global and crop-specific spectral attributions of the model trained on the ten satellite bands.

Relevance of RE-1 and SWIR-1 bands

These results highlight the relevance of RE-1 and SWIR-1 bands for crop mapping and complement the findings of earlier studies. Yi et al. [337] assessed the importance of S2 bands on the same task and found that RE-1 and SWIR-1 are more efficient in identifying crops than other bands in the Shiyang River Basin in China. Similarly, Liu et al. [191] found that RE and SWIR bands of S2 had irreplaceable effects on land cover classification.

Tab. 6.2.: Experimental results of all trained models. The best overall and crop-specific score in each experimental group is highlighted.

	All bands	Single VI							Two VIs						
All bands	✓	—	—	—	—	—	—	—	—	—	—	—	—	—	—
NDVI	—	✓	—	—	—	—	—	—	✓	✓	✓	—	—	—	—
n-NDVI	—	—	✓	—	—	—	—	—	—	—	—	—	—	✓	✓
NDRE-1	—	—	—	✓	—	—	—	—	✓	—	—	✓	✓	✓	—
NDRE-2	—	—	—	—	✓	—	—	—	—	✓	—	—	—	—	—
NDRE-3	—	—	—	—	—	✓	—	—	—	—	✓	—	—	—	—
NDMI-1	—	—	—	—	—	—	✓	—	—	—	—	✓	—	—	✓
NDMI-2	—	—	—	—	—	—	—	✓	—	—	—	—	✓	—	—
OA	0.67	0.62	0.62	0.61	0.56	0.51	**0.65**	0.63	0.64	0.62	0.62	0.67	**0.70**	0.61	0.68
macro-F1	0.67	0.63	0.62	0.61	0.57	0.52	**0.65**	0.64	0.65	0.62	0.63	0.67	**0.70**	0.62	0.69
Maize	0.65	**0.66**	0.61	0.65	0.54	0.41	0.62	0.60	0.67	0.60	0.61	0.63	**0.70**	0.61	0.66
Groundnut	0.51	0.45	0.51	0.48	0.45	0.44	**0.57**	0.50	0.49	0.51	0.57	**0.69**	0.61	0.44	0.61
Rice	0.84	0.64	0.70	0.62	0.62	0.64	0.73	**0.76**	0.66	0.74	0.66	0.77	**0.83**	0.71	0.81
Soybean	0.48	0.49	0.42	0.34	0.38	0.49	0.41	**0.50**	**0.50**	0.37	0.42	0.35	0.38	0.43	0.49
Sorghum	0.86	0.84	0.84	0.81	0.84	0.83	**0.87**	0.86	0.85	0.83	0.82	0.85	**0.88**	0.82	0.87
Yam	0.27	0.23	0.21	0.23	0.29	0.34	**0.38**	0.27	0.22	0.30	0.23	0.31	0.32	**0.33**	0.23

Tab. 6.3.: VIs used for crop mapping.

Index	Equation	Reference
NDVI	(NIR - Red) / (NIR + Red)	Rouse et al. [257]
n-NDVI	(nNIR - Red) / (nNIR + Red)	Our work
NDRE-1	(NIR - RE1) / (NIR + RE1)	Gitelson and Merzlyak [108]
NDRE-2	(NIR - RE2) / (NIR + RE2)	Our work
NDRE-3	(NIR - RE3) / (NIR + RE3)	Our work
NDMI-1	(NIR - SWIR1) / (NIR + SWIR1)	Wilson and Sader [328]
NDMI-2	(NIR - SWIR2) / (NIR + SWIR2)	Our work

6.2.3 *Enhanced Usage of Vegetation Indices*

Indices Selection In light of the insights gained from the attribution results of the satellite bands for crop mapping, we proceed with a guided selection of VIs to use as inputs, individually or in combinations of two, into our crop classification task.

Selecting RE-based indices

Given the significance of RE-1, we include the Normalized Difference Red Edge (NDRE)-1 [108] index that uses the NIR and RE-1 bands. We derive two modified indices, NDRE-2 and NDRE-3, by replacing the first RE channel with the second and third, respectively, to verify whether the relative performance of the three indices align with the attribution of their respective bands. We also incorporate the Normalized Difference Water Index (NDMI)-1, which uses the first SWIR band, and create a modified version, NDMI-2, which uses the second SWIR band, influential on rice identification. Additionally, we include the widely used NDVI, and recognizing the comparable importance of n-NIR, we introduce a modified index, narrow Normalized Difference Vegetation Index (n-NDVI), where the NIR band is replaced with n-NIR.

Selecting SWIR-based indices

Selecting NIR-based indices

Spatial resolution of selected indices

It is important to note that only the red, green, blue, and NIR bands have a resolution of 10m, while the remaining bands are originally captured either at a 20 or 60m resolution. Therefore, we ensured that all our proposed indices contain at least one of the high-resolution bands. The formula of each index is listed in Table 6.3. We retrain our model using individual indices or combinations of two indices as inputs. The results are reported in Table 6.2.3.

Modeling Results Among the models trained on a single VI, the top-performing model is based on NDMI-1, achieving an OA score of 67%. This model outperformed the baseline in identifying three crops: sorghum, groundnut, and yam. The second-best model is based on the modified version of the same index, NDMI-2, which achieved the same class accuracy as the baseline in sorghum and yam, and performed better in soybean. The third-best model, based on the NDVI, slightly outperformed the baseline on maize and soybean crops. The NDRE-3-based model achieved the lowest OA score, mainly due to its low accuracy in maize and rice crops.

Model performance using a single index

Model performance using two indices

Among the models trained on two VIs, the combination (NDRE-1, NDMI-2) achieved the highest accuracies for sorghum, maize, and rice, and outperformed the baseline in groundnut and yam crops. This combination also scored an OA score of 70%, 3pp higher than the baseline

model. The combination (NDMI-1, n-NDVI) also demonstrated comparable performance. In contrast, combining NDRE-1 and n-NDVI had the worst overall performance, mainly due to its low accuracy in rice crops, despite its higher capacity to identify yam compared to the other models. Other combinations, including (NDVI, NDRE-2) and (NDVI, NDRE-3), also displayed comparatively low overall performance.

6.2.4 Discussion

Promising results of the proposed approach

Our overall approach of exchanging the raw satellite bands with few VIs exhibits promising results. The best model based on a single index exhibited an OA 2pp lower than the baseline model, while using two indices achieved a 3pp higher accuracy in the best case. These results highlight the potential of relying solely on one or two VIs for crop identification, especially when carefully selected. In general, larger datasets benefit from increased input features, as they enable automatic learning of high-level features by the model. However, in medium-sized training datasets like ours, performance can be enhanced through careful input feature selection.

Matching attribution and VI-modeling results

As shown in Figure 6.1, SWIR-1 appears to be significantly important to identify rice and yam crops. Accordingly, we observe in Table 6.2.3 that the NDMI-1-based models achieves the best accuracy for yam and the second-best score for rice, among the single-index based models. Combining NDMI-1 with a second index also achieved high accuracies for both crops. We further observe that the proposed index NDMI-2 achieved the best accuracies on rice compared to the other single-VI based models. Additionally, it demonstrated the highest accuracies on sorghum, maize, and rice when combined with NDRE-1, outperforming all VI-based models. On the other hand, the proposed NDRE-2 and NDRE-3 indices performed poorly on the OA scores, both when used individually and when combined with NDVI. In contrast, NDRE-1 achieved high scores, particularly when combined with NDMI-1 or NDMI-2. This observation aligns with the relative average importance of the three RE bands observed in Figure 6.1, suggesting that the first band is more suitable for crop identification. Nonetheless, the second and third RE bands were of higher importance for soybean and sorghum compared to the remaining crops, which is consistent with the improvement in crop-specific accuracies achieved by the NDRE-2 and NDRE-3-based models compared to NDRE-1. In contrast, when combined with NDVI, the NDRE-1 performs better in both crops.

Inconsistencies between attribution and VI-modeling results

While the performance of the VI-based modeling aligns with the attribution results conducted on the baseline model, there were some behaviors that were not easily interpretable. For instance, according to the attribution results, soybean identification relies significantly on the first RE band, while RE-2 and RE-3 have marginal importance. Nonetheless, the soybean classification accuracy is the worst when the model is trained on the NDRE-1, while the NDRE-2 and NDRE-3 models performed better. Similarly, the RE-1 exhibits higher importance for identifying yam crop compared to the other two RE bands, but the performance of the three corresponding single-VI based models had the opposite behavior.

6.3 SUMMARY

Summary of main findings

In this chapter, we identified VIs relevant to classify each crop type, guided by the spectral attribution results of the baseline model. Our findings contribute to the growing body of evidence suggesting that the information contained within the RE and SWIR bands from the S2 multispectral satellite data is essential for discriminating crop types [218]. Based on the attribution results, we trained several models on individual and combinations of two VIs and demonstrated their ability to outperform the model trained on all bands. Importantly, the performance of these models aligned with the spectral importance in crop accuracies in most cases. Overall, our results further indicate that combining two VIs performs better than using a single index, and while some combinations improved the OA over the validation set, an examination of individual crop performance reveals that an index can be highly efficient in identifying certain crops but might struggle with others.

Limitations & future work

One limitation of our XAI-based approach is the accuracy and reliability of the baseline model. Meaningful explanation results and relevant scientific insights are conditioned by the scientific accuracy of what the model has learned during the training. Since our baseline had an OA score of 67%, we believe that further improvements in the model's performance can enhance its robustness, and consequently, the reliability of its attribution results. In future work, in addition to improving the performance of the baseline model, the dataset can be extended to cover other regions from multiple years, in order to validate our approach on a broader range of crop types and regions.

PartIV

CONCLUSION

7 Conclusion & Outlook

7.1 CONCLUSION

Availability of RS data and ML tools

In the last two decades, satellites have been launched in large numbers, and much of the Earth's surface reflectance data has become publicly available. This data explosion has coincided with advances in ML and deep networks, which are highly efficient tools for processing large datasets. However, a major challenge in this synergy is the limited interpretability of large neural networks, a characteristic that is essential for ensuring the reliability and trustworthiness of AI models deployed in sensitive EO applications.

Interpretability of DL models

Main question of the thesis

In this thesis, we address the question: *How can the interpretability of deep networks be enhanced for EO applications?* Since satellite data has characteristics different from common computer vision and tabular benchmark datasets, existing interpretability and explainability techniques in the literature require further evaluation and modification to be effectively applied to RS datasets.

Explainability Goals we addressed

In our work, we pursue two interconnected explainability goals. First, we enhance the use of XAI to justify model predictions by revealing how input features drive the final output, thereby improving the model transparency and supporting the understanding of its reasoning process. Second, we leverage these XAI-derived insights to refine feature engineering and optimize model performance, creating a feedback loop where interpretability directly contributes to the modeling enhancement.

XAI for Justification We evaluated existing interpretability tools and proposed new methods to explain DL models trained on RS data. Focusing on multimodal datasets, a common scenario in EO applications, we explored three distinct approaches to leverage different modalities for interpreting model outcomes.

Post-hoc interpretability in multimodal learning with early fusion

First, in Chapter 3.2, we adopted a straightforward modeling setup where modalities were concatenated during pre-processing after aligning their spectral and temporal dimensions. Using pixel-wise crop yield prediction as a case study, we evaluated various feature attribution techniques and found that the SVS method provided the most reliable and robust explanations. We then applied SVS to analyze model behavior under different temporal sampling scenarios, validating its reasoning against established agronomic knowledge. This validation was facilitated by mapping temporal data to crop growth stages, a standard framework in agronomy for studying and documenting yield-influencing factors.

Intrinsic interpretability in multimodal learning with intermediate fusion

Second, in Chapter 3.3, we transitioned to advanced modeling techniques for the same yield prediction task, using the intermediate fusion techniques, where each modality was processed by a separate encoder before combining their representations. After comparing various archi-

tectures, the Transformer-based model demonstrated a superior performance, balancing between model accuracy and inference speed, while offering intrinsic interpretability through attention weights. We compared intrinsic against post-hoc attribution methods and demonstrated that the AR technique produced more robust feature attributions than GA and the previously selected SVS approach. Additionally, we proposed WMA, a novel method for intrinsically estimating modality importance, and contrasted its results with SVS-derived attributions. While both methods yielded contrasting results, each presented valid arguments supporting its validity.

Multitask learning for enhanced interpretability

Finally, in Chapter 4, we exploited the inherent multimodality of RS datasets to enhance model interpretability through multitask learning. By converting certain input modalities as auxiliary prediction targets, we analyzed how errors across tasks influenced and explained the primary task's performance. The baseline was defined as conventional single-task models that relied exclusively on satellite-derived inputs (aerial, multispectral, and SAR imagery). Applied to three EO tasks, including semantic segmentation, classification, and regression, the multitask framework maintained performance comparable to the baselines while reducing dependency on additional input data at deployment (since the model now predicts these modalities rather than requiring them as inputs). Moreover, we demonstrated that error propagation patterns across tasks exhibited strong correlations, providing a novel mechanism for explaining predictions in the main task.

XAI for Improvement After having explored the various methodologies to justify the outcome of EO models, we leveraged the corresponding explanation results to improve the feature engineering step, ultimately leading to an improvement in the model performance. To fulfill this goal, we explored two directions.

XAI-guided optimization of the input space

First, guided by the feature attribution results, we adopted a cyclic approach in Chapter 5 to iteratively reduce the input dimensionality. The process begins by training the model on all modalities, including all spectral bands and time steps, followed by explaining its predictions using six distinct attribution methods to rank features by their importance. The least important features along the spectral and temporal dimensions are then removed, and the model is retrained on the updated dataset. We repeated this cycle across three EO datasets, yielding surprising results: a significant proportion of features could be removed without substantial model accuracy degradation. Among the six attribution methods evaluated, the SVS method provided the most robust estimates, aligning with findings from Chapter 3.2. Further experiments identified the necessary features for each task, defined as those whose removal caused a pronounced performance drop. Interpreting these results in the context of each dataset offered valuable insights into the utility of specific modalities and time steps, deepening our understanding of their roles in model performance.

XAI-guided selection of VIs

Second, we focused in Chapter 6 on the specific case of VIs, which are commonly used in EO applications, particularly for agricultural tasks. Given the extensive number of indices available in the literature, we

proposed an XAI-guided approach to select the most suitable VIs for the target task. In this study, conducted on a crop classification task, we first trained a model using all 12 bands of the multispectral satellite data. We then explained the model's predictions using the SVS method to identify the most important bands for each class. These attribution results guided the selection of relevant VIs, after which we retrained the model by replacing the original 12 bands with either one or two indices, leading to successful outcomes: the results demonstrated that a single VI achieved performance comparable to the baseline (all 12 bands), while using two VIs surpassed it. Together with the cyclic approach described earlier, this study confirms the potential of explanation methods to enhance feature engineering and modeling in EO tasks.

7.2 CHALLENGES & FUTURE WORK

Following the comprehensive summary in Section 7.1, we conclude this thesis by outlining three high-priority research directions to extend the presented work.

Explanation consistency across models Throughout our experiments, interpretability analysis was conducted on individual model instances per dataset. This raises an important question about the consistency of explanations when alternative models are considered, whether trained with different randomization seeds, data splits, or architectural configurations. For instance, in Chapter 3, Section 3.2, models were trained using 10-fold cross-validation, yet explanations were derived solely from the best-performing fold. Future work could systematically evaluate explanation variability across all folds and assess its impact on agronomic validation of the explanation results. Similarly, in Section 3.3, multiple intermediate fusion architectures were evaluated, and within each architecture we tested various hyperparameter configurations. Here as well, the explanation analysis could be repeated on each configuration architecture to quantify result variance and its implications.

Experiments which require an explanation consistency check

Overall, XAI results are inherently model-dependent, as they explain the specific reasoning learned by a given model, with no guarantee of generalizability to other model variants. Nevertheless, expectations are that models trained on the same dataset for the same task would learn similar reasoning patterns, which translates as expectations that the explanation results should maintain some similarity across models. Quantifying the robustness of explanations across model variants would provide empirical verification for this hypothesis.

Explanations are model-instance specific

Rigorous evaluation of proposed methods In the literature, the development of new explanation methods requires rigorous qualitative and quantitative evaluation to establish their generalizability and robustness. While this thesis has introduced several novel explanation frameworks and techniques, additional evaluations would further validate their efficiency. For instance, in Chapter 3, Section 3.3, we proposed the WMA method for estimating modality importance. However, comparisons with the established post-hoc SVS approach revealed divergent results. In order to rank and choose between both techniques, future work should rely on quantitative evaluations. Namely, the sensitivity and infidelity metric used to evaluate feature attribution can be adjusted and modified to also systematically assess and compare these modality attribution techniques.

Evaluation metrics for quantitative assessment

Furthermore, applying both WMA and SVS to other EO tasks could reveal whether their results converge in different application contexts. The multitask learning framework presented in Chapter 4 successfully demonstrated consistent explanatory power across datasets in analyzing the main task's predictions. Similarly, the XAI-guided input space optimization in Chapter 5 showed consistent efficiency when applied to three distinct EO datasets. For the VIs selection framework developed in Chapter 6, valuable extensions could include application to water-

Evaluation across datasets and EO tasks

and drought-related tasks. This would enable optimization of water and drought indices selection, potentially revealing the efficacy and sufficiency of such indices for related applications, including surface water mapping, flood detection, and agricultural drought monitoring.

Interdisciplinary studies for scientific insights Beyond its roles in model justification and performance improvement, XAI can further serve the purpose of scientific discovery. In particular, when uncovering the reasoning process of the model, domain experts can validate this reasoning against established knowledge. In fact, explanations can yield two possible outcomes. First, when the model's reasoning aligns with domain knowledge, it confirms that the model has learned meaningful input-target relationships, which increases confidence in its generalization capability to new input samples. Second, when discrepancies emerge between the model's reasoning and existing knowledge, this does not automatically indicate spurious correlations. Instead, it may suggest previously undiscovered patterns that call for expert investigation. In such cases, domain specialists can step in and conduct follow-up studies to assess the scientific validity of these novel patterns.

Alignment between explanations and domain knowledge

Therefore, interdisciplinary collaborations play an important role in verifying patterns which lack support in existing literature. A concrete example emerges from Chapter 3, Section 3.3, where we explained the task of crop yield prediction, and where our attribution analysis revealed some discrepancies between the model's focus and literature-established critical growth stages for crop yield. While certain growth stages well-documented in agronomic studies appeared less influential in our model's decisions, other time steps emerged as significant predictors. These findings invite agronomists to examine whether the identified influential periods represent overlooked yield determinants or require model refinement.

Experts can verify novel patterns

PartV

APPENDIX

A

Explainability for Earth Observation

In the following, we describe the search query used to create Figure 2.1, in which we compare the increase in number of publications in ML and XAI, in EO. In Scopus [1] database, the search query used consists of two major parts: keywords related to EO, and keywords related to ML and XAI. All the nested keywords are connected via the `OR` operator, while the two parts are connected via the `AND` operator. Due to the interchangeably used taxonomy in each fields, we added additional keywords to the generally known terms to optimize the true positives rate.

Query 1: Machine Learning in Earth Observation:

```
[ Earth observation OR remote sensing OR
((satellite OR aerial OR airborne OR spaceborne OR radar)
AND (image OR data))
OR LiDAR OR SAR OR UAV OR Sentinel OR Landsat OR MODIS]
AND
[ deep learning OR machine learning OR artificial intelligence ]
```

Query 2: Explainable AI in Earth Observation:

```
[ Earth observation OR remote sensing OR
((satellite OR aerial OR airborne OR spaceborne OR radar)
AND (image OR data))
OR LiDAR OR SAR OR UAV OR Sentinel OR Landsat OR MODIS]
AND
[ deep learning OR machine learning OR artificial intelligence ]
AND
[ xai OR feature importance OR SHAP ]
```

In addition to the keywords listed above, four further criteria were applied to both queries: the time range (2017–2024), language restriction (English), document types (journal articles and conference papers), and relevant research fields, including Earth and Planetary Sciences, Environmental Science, Agriculture and Biological Sciences, Computer Science, and Engineering.

[1] https://www.scopus.com/

Evaluation Metrics

B

Regression tasks Models designed for regression tasks are optimized using the MAE as a loss function. It is also employed for model evaluation, alongside the RMSE and R^2 metrics. The respective formulas for each metric are as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Classification tasks When dealing with either binary or multi-class classification problem, commonly used evaluation metrics are the (overall) accuracy, micro-F1, micro-precision, micro-recall, macro-F1, macro-precision, and macro-recall scores.

$$\text{Overall Accuracy} = \frac{\sum_{\text{instances}} \mathbb{I}(\text{Predicted Class} = \text{Actual Class})}{\text{Total Instances}},$$

$$\text{Micro Precision} = \frac{\sum_{\text{classes}} \text{TP}_{\text{class}}}{\sum_{\text{classes}} (\text{TP}_{\text{class}} + \text{FP}_{\text{class}})},$$

$$\text{Micro Recall} = \frac{\sum_{\text{classes}} \text{TP}_{\text{class}}}{\sum_{\text{classes}} (\text{TP}_{\text{class}} + \text{FN}_{\text{class}})},$$

$$\text{Micro F1} = 2 \times \frac{\text{Micro Precision} \times \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}},$$

$$\text{Macro Precision} = \frac{\sum_{c=1}^{C} p^c}{C},$$

$$\text{Macro Recall} = \frac{\sum_{c=1}^{C} r^c}{C},$$

$$\text{Macro F1} = \frac{\sum_{c=1}^{C} f_1^c}{C},$$

where TP stands for True Positive, FP for False Positive, FN for False Negative, TN for True Negative, $\mathbb{I}$ is an indicator function that returns 1 if the condition is true and 0 otherwise, C is the number of classes, $\{p^1, p^2, \ldots, p^C\}$ are class-wise precisions, $\{r^1, r^2, \ldots, r^C\}$ are class-wise recalls, and $\{f_1^1, f_1^2, \ldots, f_1^C\}$ are class-wise F1-scores.

Segmentation tasks Segmentation models are evaluated using primarily two metrics: accuracy and IoU. Accuracy measures the proportion of correctly classified pixels out of all pixels in the image. For semantic segmentation tasks, accuracy can be misleading because it doesn't differentiate between false positives (pixels incorrectly labeled as belonging to a class) and false negatives (pixels correctly belonging to a class but misclassified). In contrast, IoU complements the accuracy metric by comparing the overlap between the predicted mask and the ground truth mask:

$$\text{Accuracy} = \frac{\text{Number of Correctly Classified Pixels}}{\text{Total Number of Pixels}},$$

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Total Area}}.$$

C

Vegetation Indices & Growth Stages

C.1 VEGETATION INDICES

For the experiment exchanging the satellite bands with vegetation indices, we selected 10 indices commonly used for crop monitoring, as summarized in Table C.2

C.2 GROWTH STAGES

Tab. C.1.: Growth stages adapted for each crop.

Rapeseed [173]	**Wheat** [173]	**Soybean** [209]
Bud development	Bud development	Emergence
Leaf development	Leaf development	Unifoliolate
Shoot development	Tillering	1st-25th Trifoliolate
Heading	Shoot development	Beginning Bloom
Flowering	Bolting	Full Bloom
Development of fruit	Heading	Beginning Pod
Ripening	Flowering	Full Pod
Senescence	Development of fruit	Beginning Seed
	Ripening	Full Seed
	Senescence	Beginning Maturity
		Full Maturity

To aggregate or visualize the temporal attributions by growth stages, we use the crop-specific scales described in Table C.1.

Phenology data was provided by xarvio [1], using in-house developed and commercially deployed growth stage models that estimate cultivar-specific growth stages of different crops on a daily base. These models are trained (as a ML model) per country and crop, based on local cultivar-specific growth stage observations acquired in field trials (among other observations) as well as additional data sources such as weather data, to best account for local growth conditions.

[1] www.xarvio.com, accessed 13 January 2025.

Tab. C.2.: Vegetation indices used in the yield modeling and their corresponding equations. For simplification purposes, certain satellite bands are replaced as follows: B for B02, G for B03, R for B04, RE1 for B05, RE2 for B06, and N for B08.

Vegetation Index	Formula	Reference
Chlorophyll Index Green (CIG)	(N / G) - 1.0	[110]
Chlorophyll Index Red Edge (CIRE)	(N / RE1) - 1	[110]
Green Normalized Difference Vegetation Index (GNDVI)	(N - G)/(N + G)	[111]
Normalized Difference Vegetation Index (NDVI)	(N - R)/(N + R)	[257]
Normalized Difference Vegetation Index (NDYI)	(G - B) / (G + B)	[288]
Ratio Vegetation Index RVI	RE2 / R	[32]
Wide Dynamic Range Vegetation Index (WDRVI)	(0.1 * N - R) / (0.1 * N + R)	[109]
Normalized Green Red Difference Index (NGRDI)	(G - R) / (G + R)	[311]
Modified Chlorophyll Absorption Ratio Index / Optimized Soil-Adjusted Vegetation Index (MCARI/OSAVI)	(((RE2 - RE1) - 0.2 * (RE2 - G)) * (RE2 / RE1)) / (1.16 * (RE2 - RE1) / (RE2 + RE1 + 0.16))	[330]
Transformed Chlorophyll Absorption Ratio Index / Optimized Soil-Adjusted Vegetation Index (TCARI/OSAVI)	(3 * ((RE2 - RE1) - 0.2 * (RE2 - G) * (RE2 / RE1))) / (1.16 * (RE2 - RE1) / (RE2 + RE1 + 0.16))	[330]

D Multimodal learning via early fusion

D.1 PADDED BASELINE IN ARG-S

We report the results of the evaluation of different XAI methods using the padded baseline in Figures D.1 and D.2 for qualitative evaluation, and in Figure D.3 for qualitative evaluation.

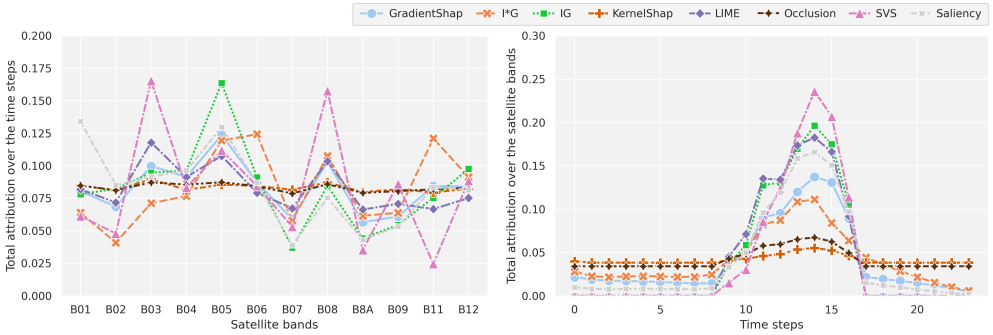


Fig. D.1.: XAI Methods results using the padded baseline and ARG-S dataset. On the left is the spectral importance and on the right the temporal importance.

D.2 QUANTITATIVE EVALUATION IN OTHER DATASETS

The evaluation results of the XAI methods are consistent across the different crop yield datasets: SVS exhibits higher robustness compared to the other attribution methods, according to the infidelity and sensitivity scores. The results are displayed in Figure D.4.

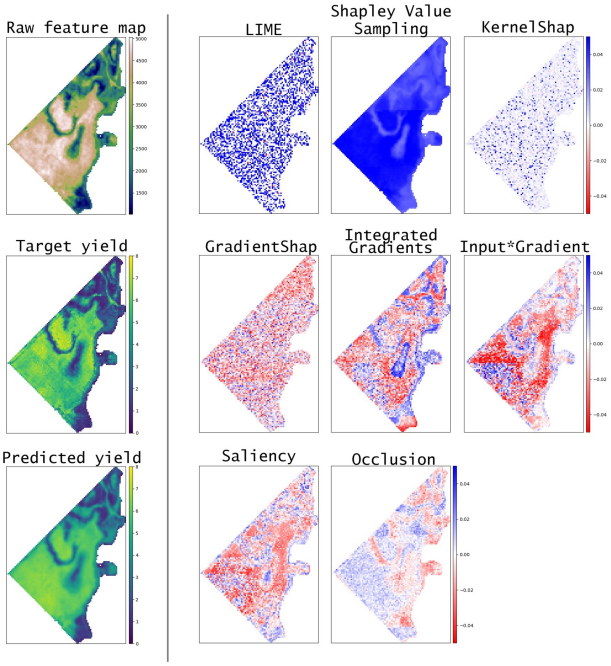


Fig. D.2.: Qualitative evaluation of the feature attribution methods using the padded baseline on ARG-S dataset: Attribution maps of band B08 of time step 14 (March of the second year).

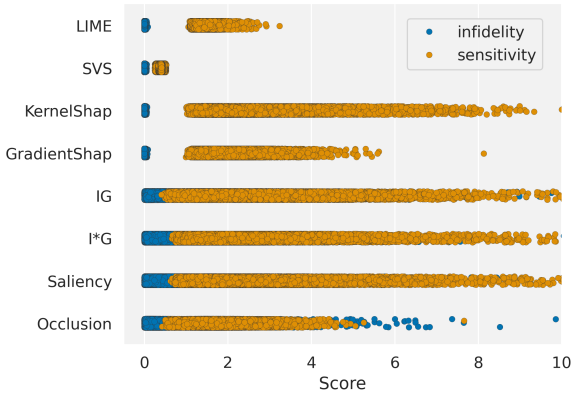


Fig. D.3.: Quantitative evaluation of the feature attribution methods using the padded baseline on ARG-S dataset.

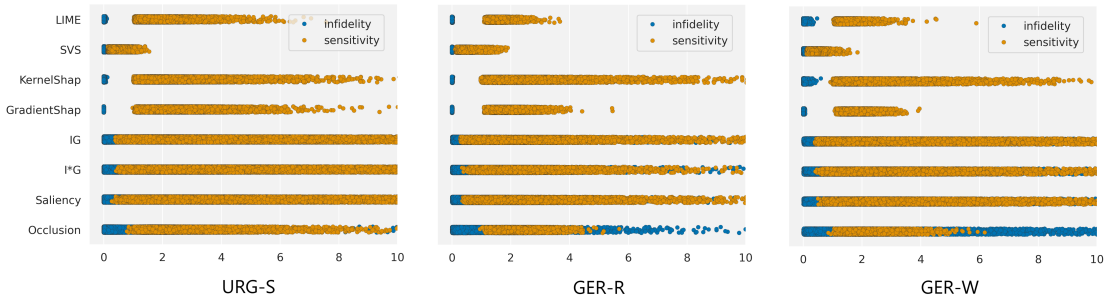


Fig. D.4.: Quantitative evaluation of the feature attribution methods using the mean baseline on URG-S, GER-R, and GER-W datasets.

E

Multimodal learning via late fusion

E.1 ATTENTION WEIGHTS DISTRIBUTION

Figure E.1 displays the comparison of attention weights distribution across different layers of the Transformer encoder of satellite and weather modalities, for three random corn fields.

E.2 TEMPORAL ATTRIBUTIONS

a. Additional corn fields In Figure E.2, we display the temporal attributions of the three evaluated methods for the same three random corn fields presented in E.1.

b. Quantitative Evaluation We evaluate the methods for estimating temporal attributions using the infidelity and sensitivity scores on soybean, wheat and rapeseed fields in Argentina and Germany. The results displayed in Figure E.3 reveal that the infidelity scores are similarly low in all the datasets, while the sensitivity scores reflect varying ranks. AR exhibits lowest sensitivity across all fields, except for soybean crops in Argentina, while GA achieves comparable score as SVS in wheat fields, and a worse performance in rapeseed and soybean fields. Overall, a comparison of the two intrinsic methods reveals that AR consistently provides more stable attributions compared to GA.

c. Agronomic validation To illustrate how the temporal attributions can be interpreted in the light of agronomically meaningful periods, we retrieve some soybean fields for which we obtained approximated information about the start and end dates of different soybean growth stages. We overlap the phenological periods on the temporal attribution plots for three random soybean fields in Figure 3.31. The results display distinctive patterns across the modalities. We first analyze the satellite encoder:

- Satellite data has a relatively low importance during the initial two stages of **emergence** and **leaf development** (i.e. unifoliolate and trifoliolates). During this vegetation phase, the main stem nodes and their branching are developing, influencing the canopy structure and the final number of nodes [170]. As a result, a poor canopy expansion might be identifiable in the satellite pixels, and thus used by the model to correlate with lower yield values. However, the results suggest that the model does not rely significantly on this cue.

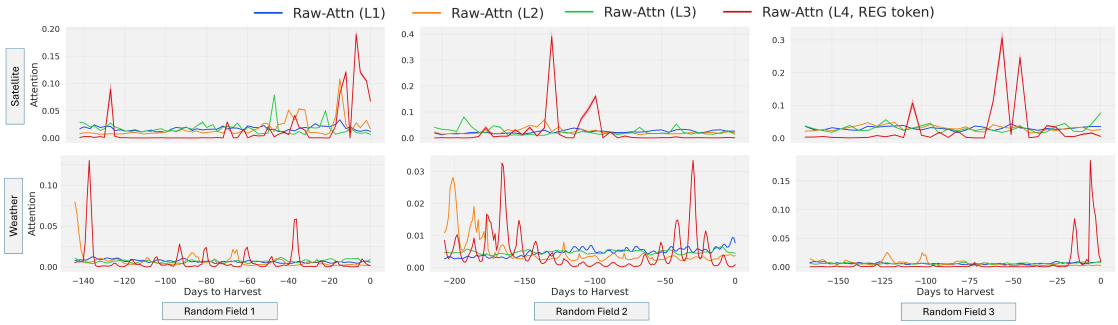


Fig. E.1.: Total attention weights attending at each time step for the first three attention layers, and the regression token weights in the final layer. The results are averaged across 32 randomly selected pixels from three random fields, and are displayed for the satellite (a) and weather (b) Transformer encoders. The light buffer regions represent the 95% confidence interval around the average value.

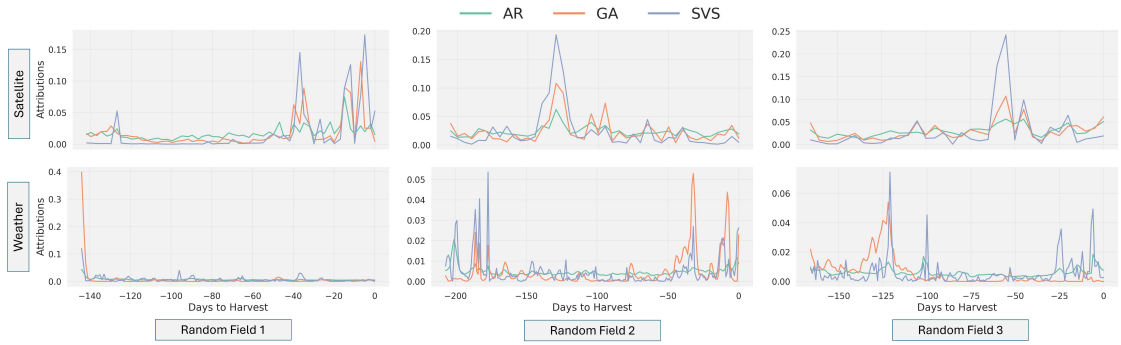


Fig. E.2.: Field-level average attributions of the satellite and weather modalities, for three random corn fields.

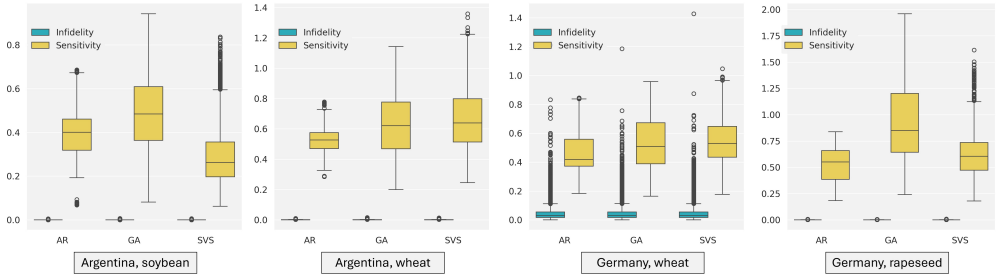


Fig. E.3.: Infidelity and Sensitivity scores of the temporal attributions estimated by AR, GA, and SVS methods, for soybean and wheat fields from Argentina, and wheat and rapeseed fields from Germany.

- We then observe that attribution values slightly increase at the **blooming** and **podding stages**, particularly in the field **c**. Defoliation of the plant during late blooming is, in fact, known to negatively affect yield [209]. The total number of mature nodes and pods that develop during these two stages is also correlated

with yield and may alter the field's landscape, thereby serving as an early indicator of potential yield.

- In contrast, the model seems to rely more on the **seeding stage** in fields **a** and **b**. In fact, leaf loss of 100% has been shown to reduce yields by 80% during early seeding [209]. Another visual hint that can be captured by the model through the satellite data is the green seeds, which appear only during this stage.

- The final **maturity stage** has a moderate level of importance across the three fields. This can be explained by the identifiability of the pods: 95% of the pods on the main stem reach their mature pod color at this stage, which can potentially serve as a visual cue of the yield [170].

Weather data attributions exhibits different patterns, as it provides a different type of information to the model. An examination of 50 additional soybean fields revealed that the climate conditions during the early growing period and as the harvesting date approaches typically have the greatest influence on the model's predictions. More particularly:

- During **emergence** stage, emergence speed is impacted by temperature and moisture conditions [209]. In addition to soybeans being highly sensitive to salt, soil composition plays a critical role in nitrogen fixation and nutrient absorption, processes that occur during the early growing stages. Meanwhile, weather conditions are essential in regulating nutrient dynamics in the soil, affecting both nutrient availability and uptake by plants [53, 77, 307].

- The **podding** is the most crucial period for seed yield, and any stress from **late podding** until **full seeding** causes more yield reduction than at any other time [209]. Thus, if the weather conditions are unfavorable, it is usually recommended to compensate with appropriate irrigation strategies. However, the model does not appear to have caught important patterns connecting weather conditions during podding and seeding to the predicted yield.

- While stress during the **maturity stage** has almost no effect on yield, adequate weather conditions are required for the soybeans to dry and reach at least 15% moisture to be ready for harvest [209]. High drought might, however, cause significant losses. This weather potential influence might explain the high attribution values observed at this stage.

Weather patterns remained consistent in most of the 50 soybean fields we examined. However, the temporal importance of satellite data showed more variance, predominantly fluctuating between the podding and seeding stages. Overall, this analysis highlights how some of the time steps that the model focuses on within each modality align with their corresponding agronomic significance, while other important patterns appear to be overlooked by the model. These observations, however, require further careful verification in collaboration with agronomy experts.

Additional results Figure E.4 compares WMA and SVS scores for 50 random fields of soybean and wheat crops in Argentina, soybean in Uruguay, and wheat and rapeseed in Germany. Consistent with the findings for corn fields shown in Figure 3.34, we observe that satellite data is the most influential modality according to Shapley-based scores, with terrain and soil having a marginal contribution. In contrast, WMA scores suggest a reduced influence of satellite data, in favor of soil and weather modalities. Terrain elevation properties show minimal significance to the model across both interpretability techniques.

When comparing crops across different regions, the modality scores for soybean fields are consistent between Argentina and Uruguay, with weather being the most influential modality, followed by satellite and then soil. In contrast, models trained on wheat crops demonstrate a stronger reliance on soil and reduced usage of the satellite data in Germany compared to Argentina. Overall, these regional differences likely reflect the impact of climate conditions, given that fields in Argentina and Uruguay are located in nearby regions and share similar climates, whereas German wheat crops grow in a different climatic environment than Argentinian wheat fields. Additionally, the climate impacts satellite data availability; for example, frequent cloudy weather in some regions renders several satellite images unusable for the model, further influencing the modality importance.

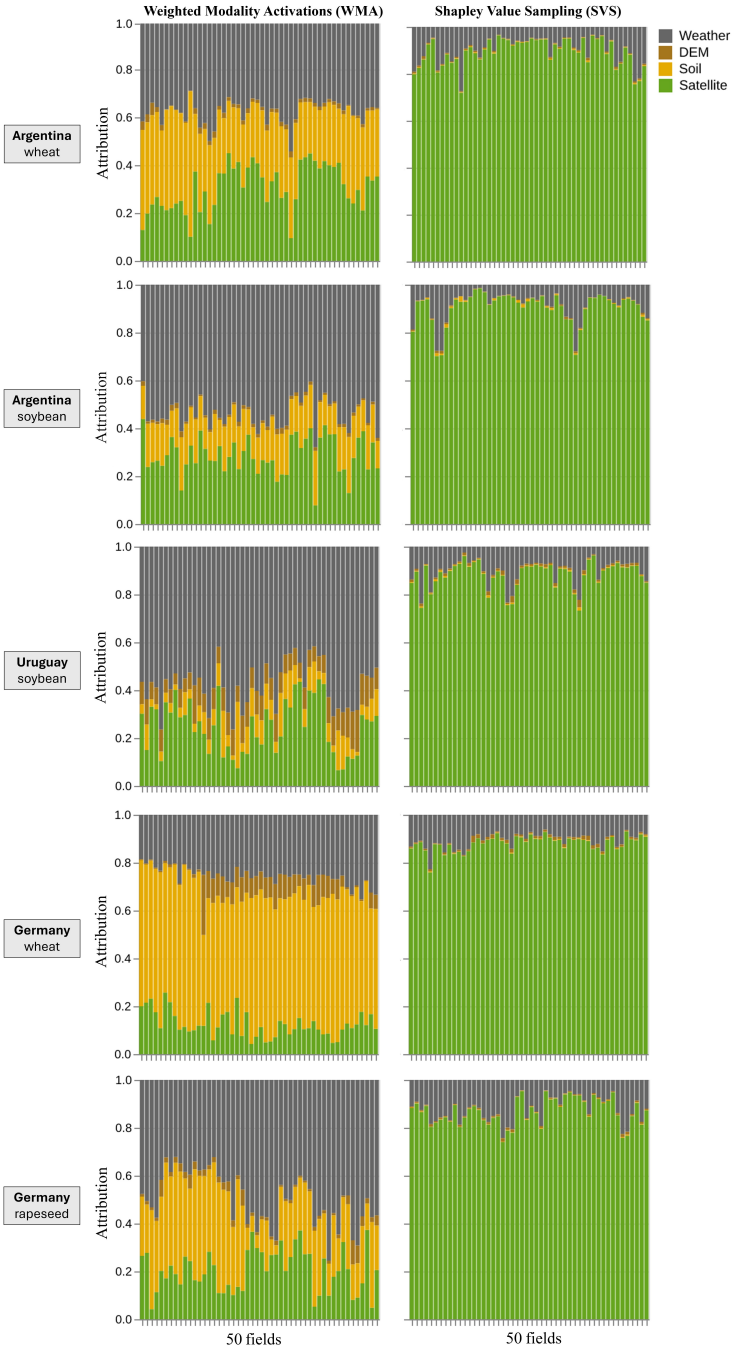


Fig. E.4.: Comparing the modality importance using WMA and SVS scores for 50 fields of wheat and soybean in Argentina, soybean in Uruguay, and Wheat and rapeseed in Germany.

Bibliography

[1]Samira Abnar and Willem Zuidema. "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4190–4197 (cit. on pp. 65, 69).

[2]Edmundo Acevedo, Paola Silva, and Herman Silva. "Growth and wheat physiology, development". In: *Laboratory of Soil-Plant-Water Relations. Faculty of Agronomy and Forestry Sciences. University of Chile. Casilla* 1004 (2006) (cit. on pp. 49, 50).

[3]Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160 (cit. on pp. 17, 19).

[4]Julius Adebayo, Justin Gilmer, Michael Muelly, et al. "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 23, 25, 109).

[5]Steve Ahlswede, Christian Schulz, Christiano Gava, et al. "TreeSatAI Benchmark Archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing". In: *Earth System Science Data Discussions* 2022 (2022), pp. 1–22 (cit. on pp. 86, 94).

[6]Ata Akbari Asanjan, Tiantian Yang, Kuolin Hsu, et al. "Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks". In: *Journal of Geophysical Research: Atmospheres* 123.22 (2018), pp. 12–543 (cit. on p. 15).

[7]Husam AH Al-Najjar, Biswajeet Pradhan, Ghassan Beydoun, et al. "A novel method using explainable artificial intelligence (XAI)-based Shapley Additive Explanations for spatial landslide prediction using Time-Series SAR dataset". In: *Gondwana Research* 123 (2023), pp. 107–124 (cit. on p. 24).

[8]Guillaume Alain and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes". In: *arXiv preprint arXiv:1610.01644* (2016) (cit. on p. 67).

[9]Mouad Alami Machichi, loubna El mansouri, Yasmina Imani, et al. "Crop mapping using supervised machine learning and deep learning: a systematic literature review". In: *International Journal of Remote Sensing* 44.8 (2023), pp. 2717–2753 (cit. on p. 12).

[10]David Alvarez Melis and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks". In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 25, 43).

[11]Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. "Towards better understanding of gradient-based attribution methods for Deep Neural Networks". In: *International Conference on Learning Representations*. 2018 (cit. on p. 31).

[12]Giuseppina Andresini, Annalisa Appice, and Donato Malerba. "SILVIA: An eXplainable Framework to Map Bark Beetle Infestation in Sentinel-2 Images". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2023) (cit. on p. 24).

[13] Jason Ansel, Edward Yang, Horace He, et al. "Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation". In: *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 2024, pp. 929–947 (cit. on p. 31).

[14] Daniel W Apley and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.4 (2020), pp. 1059–1086 (cit. on p. 21).

[15] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. "Gated multimodal units for information fusion". In: *arXiv preprint arXiv:1702.01992* (2017) (cit. on p. 16).

[16] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "Explaining Predictions of Non-Linear Classifiers in NLP". In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. 2016, pp. 1–7 (cit. on p. 22).

[17] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "Explaining Recurrent Neural Network Predictions in Sentiment Analysis". In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2017, pp. 159–168 (cit. on p. 22).

[18] Peter M Atkinson and Adrian RL Tatnall. "Introduction neural networks in remote sensing". In: *International Journal of remote sensing* 18.4 (1997), pp. 699–709 (cit. on p. 13).

[19] Halit Enes Aydin and Muzaffer Can Iban. "Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations". In: *Natural Hazards* 116.3 (2023), pp. 2957–2991 (cit. on p. 24).

[20] MR Azimi-Sadjadi and SA Zekavat. "Cloud classification using support vector machines". In: *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No. 00CH37120)*. Vol. 2. IEEE. 2000, pp. 669–671 (cit. on p. 13).

[21] Sebastian Bach, Alexander Binder, Grégoire Montavon, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140 (cit. on p. 22).

[22] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. "End-to-end attention-based large vocabulary speech recognition". In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2016, pp. 4945–4949 (cit. on p. 61).

[23] Jean-Stéphane Bailly, M Arnaud, and C Puech. "Boosting: A classification method for remote sensing". In: *International Journal of Remote Sensing* 28.7 (2007), pp. 1687–1710 (cit. on p. 13).

[24] Amit Banerjee, Philippe Burlina, and Chris Diehl. "A support vector method for anomaly detection in hyperspectral imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 44.8 (2006), pp. 2282–2291 (cit. on p. 13).

[25] Richard L Bankert. "Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network". In: *Journal of Applied Meteorology and climatology* 33.8 (1994), pp. 909–918 (cit. on p. 13).

[26]Teo Beker, Homa Ansari, Sina Montazeri, Qian Song, and Xiao Xiang Zhu. "Deep learning for subtle volcanic deformation detection with InSAR data in central volcanic zone". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–20 (cit. on p. 20).

[27]Mariana Belgiu and Ovidiu Csillik. "Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis". In: *Remote sensing of environment* 204 (2018), pp. 509–523 (cit. on p. 14).

[28]Hans Georg Beyer, Claudio Costanzo, and Detlev Heinemann. "Modifications of the Heliosat procedure for irradiance estimates from satellite images". In: *Solar Energy* 56.3 (1996), pp. 207–212 (cit. on p. 13).

[29]Umang Bhatt, Adrian Weller, and José MF Moura. "Evaluating and aggregating feature-based model explanations". In: *arXiv preprint arXiv:2005.00631* (2020) (cit. on p. 43).

[30]Umang Bhatt, Adrian Weller, and José MF Moura. "Evaluating and aggregating feature-based model explanations". In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 3016–3022 (cit. on p. 19).

[31]Or Biran and Courtenay Cotton. "Explanation and justification in machine learning: A survey". In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. 1. 2017, pp. 8–13 (cit. on p. 16).

[32]Gerald S Birth and George R McVey. "Measuring the color of growing turf with a reflectance spectrophotometer 1". In: *Agronomy Journal* 60.6 (1968), pp. 640–643 (cit. on p. 142).

[33]Katalin Blix, Gustau Camps-Valls, and Robert Jenssen. "Gaussian process sensitivity analysis for oceanic chlorophyll estimation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.4 (2017), pp. 1265–1277 (cit. on p. 20).

[34]Maurice Borgeaud, Noémy Scheidegger, Muriel Noca, et al. "SwissCube: the first entirely-built Swiss student satellite with an Earth observation payload". In: *Small Satellite Missions for Earth Observation: New Developments and Trends*. Springer. 2010, pp. 207–213 (cit. on p. 11).

[35]Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32 (cit. on p. 21).

[36]Gunnar Jakob Briem, Jon Atli Benediktsson, and Johannes R Sveinsson. "Boosting, bagging, and consensus based classification of multisource remote sensing data". In: *Multiple Classifier Systems: Second International Workshop, MCS 2001 Cambridge, UK, July 2–4, 2001 Proceedings 2*. Springer. 2001, pp. 279–288 (cit. on p. 13).

[37]Martin Brown, Steve R Gunn, and Hugh G Lewis. "Support vector machines for optimal classification and spectral unmixing". In: *Ecological Modelling* 120.2-3 (1999), pp. 167–179 (cit. on p. 13).

[38]Martin Brown, Hugh G Lewis, and Steve R Gunn. "Linear spectral mixture models and support vector machines for remote sensing". In: *IEEE Transactions on geoscience and remote sensing* 38.5 (2000), pp. 2346–2360 (cit. on p. 13).

[39]Manuel Campos-Taberner, Francisco Javier García-Haro, Beatriz Martínez, et al. "Understanding deep learning in land use classification based on Sentinel-2 time series". In: *Scientific reports* 10.1 (2020), p. 17188 (cit. on p. 20).

[40] Jonathan Cheung-Wai Chan, Chengquan Huang, and Ruth Defries. "Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting". In: *IEEE Transactions on Geoscience and Remote Sensing* 39.3 (2001), pp. 693–695 (cit. on p. 13).

[41] Jonathan Cheung-Wai Chan and Desiré Paelinckx. "Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery". In: *Remote Sensing of Environment* 112.6 (2008), pp. 2999–3011 (cit. on p. 13).

[42] Hila Chefer, Shir Gur, and Lior Wolf. "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 397–406 (cit. on pp. 65, 69).

[43] Guanyu Chen, Peng Jiao, Qing Hu, Linjie Xiao, and Zijian Ye. "SwinSTFM: Remote sensing spatiotemporal fusion using Swin transformer". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–18 (cit. on p. 15).

[44] Hao Chen, Zipeng Qi, and Zhenwei Shi. "Remote sensing image change detection with transformers". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–14 (cit. on p. 15).

[45] Jia Chen, Lizhe Wang, Ruyi Feng, et al. "CycleGAN-STF: Spatiotemporal fusion via CycleGAN-based image generation". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.7 (2020), pp. 5851–5865 (cit. on p. 15).

[46] Song Chen. *PM2.5 Data of Five Chinese Cities*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52K58. 2017 (cit. on p. 108).

[47] Yushi Chen, Chunyang Li, Pedram Ghamisi, Xiuping Jia, and Yanfeng Gu. "Deep fusion of remote sensing data for accurate classification". In: *IEEE Geoscience and Remote Sensing Letters* 14.8 (2017), pp. 1253–1257 (cit. on p. 16).

[48] Dongcai Cheng, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. "FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.12 (2017), pp. 5769–5783 (cit. on p. 15).

[49] Gong Cheng, Peicheng Zhou, and Junwei Han. "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images". In: *IEEE transactions on geoscience and remote sensing* 54.12 (2016), pp. 7405–7415 (cit. on p. 15).

[50] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *arXiv preprint arXiv:1412.3555* (2014) (cit. on pp. 109, 121).

[51] Martin Claverie, Eric F Vermote, Belen Franch, and Jeffrey G Masek. "Evaluation of the Landsat-5 TM and Landsat-7 ETM+ surface reflectance products". In: *Remote Sensing of Environment* 169 (2015), pp. 390–403 (cit. on p. 14).

[52] Miruna A Clinciu and Helen F Hastie. "A survey of explainable AI terminology". In: *1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence 2019*. Association for Computational Linguistics. 2019, pp. 8–13 (cit. on pp. 16, 18).

[53]Dennis L Corwin. "Climate change impacts on soil salinity in agricultural areas". In: *European Journal of Soil Science* 72.2 (2021), pp. 842–862 (cit. on p. 149).

[54]LJ Crone, LM McMillin, and DS Crosby. "Constrained regression in satellite meteorology". In: *Journal of Applied Meteorology and Climatology* 35.11 (1996), pp. 2023–2035 (cit. on p. 13).

[55]Chaoya Dang, Ying Liu, Hui Yue, JiaXin Qian, and Rong Zhu. "Autumn crop yield prediction using data-driven approaches:-support vector machines, random forest, and deep neural network methods". In: *Canadian journal of remote sensing* 47.2 (2021), pp. 162–181 (cit. on p. 45).

[56]RS De Fries, M Hansen, JRG Townshend, and R Sohlberg. "Global land cover classifications at 8 km spatial resolution: The use of training data derived from Landsat imagery in decision tree classifiers". In: *International Journal of Remote Sensing* 19.16 (1998), pp. 3141–3168 (cit. on p. 13).

[57]Ruth DeFries, Matthew Hansen, Marc Steininger, et al. "Subpixel forest cover in central Africa from multisensor, multitemporal data". In: *Remote Sensing of Environment* 60.3 (1997), pp. 228–246 (cit. on p. 13).

[58]Zhipeng Deng, Hao Sun, Shilin Zhou, et al. "Multi-scale object detection in remote sensing imagery with convolutional neural networks". In: *ISPRS journal of photogrammetry and remote sensing* 145 (2018), pp. 3–22 (cit. on p. 15).

[59]Johann Desloires, Dino Ienco, and Antoine Botrel. "Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods". In: *Computers and Electronics in Agriculture* 209 (2023), p. 107807 (cit. on p. 14).

[60]Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on p. 61).

[61]Yonas B Dibike, Slavco Velickov, and Dimitri Solomatine. "Support vector machines: Review and applications in civil engineering". In: *Proceedings of the 2nd Joint Workshop on Application of AI in Civil Engineering*. 2000, pp. 215–218 (cit. on p. 13).

[62]Wulf A Diepenbrock. "Yield analysis of winter oilseed rape (Brassica napus L.): a review". In: *Field crops research* 67.1 (2000), pp. 35–49 (cit. on p. 49).

[63]Daisy Yi Ding, Chloé Simpson, Stephen Pfohl, et al. "The effectiveness of multitask learning for phenotyping with electronic health records data". In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. Vol. 24. NIH Public Access. 2019, p. 18 (cit. on pp. 83, 102).

[64]Lei Ding, Jing Zhang, Haitao Guo, et al. "Joint spatio-temporal modeling for semantic change detection in remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* (2024) (cit. on p. 15).

[65]Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. "Visualizing and understanding neural machine translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 1150–1159 (cit. on p. 22).

[66]Wenqian Dong, Yufei Yang, Jiahui Qu, Weiying Xie, and Yunsong Li. "Fusion of hyperspectral and panchromatic images using generative adversarial network and image segmentation". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–13 (cit. on p. 15).

[67] Finale Doshi-Velez and Been Kim. "Considerations for evaluation and generalization in interpretable machine learning". In: *Explainable and interpretable models in computer vision and machine learning* (2018), pp. 3–17 (cit. on p. 25).

[68] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017) (cit. on pp. 17, 19).

[69] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR* (2021) (cit. on p. 61).

[70] I. D. Downey, C. H. Power, I. Kanellopoulos, and G. Wilkinson. "A performance comparison of Landsat TM land cover classification based on neural network techniques and traditional maximum likelihood and minimum distance algorithms." In: (1992), 11 pp. (Cit. on p. 13).

[71] Bo Du and Liangpei Zhang. "Random-selection-based anomaly detector for hyperspectral imagery". In: *IEEE Transactions on Geoscience and Remote sensing* 49.5 (2010), pp. 1578–1589 (cit. on p. 13).

[72] Rudresh Dwivedi, Devam Dave, Het Naik, et al. "Explainable AI (XAI): Core ideas, techniques, and solutions". In: *ACM Computing Surveys* 55.9 (2023), pp. 1–33 (cit. on p. 17).

[73] Jessica Echterhoff, An Yan, Kyungtae Han, et al. "Driving through the Concept Gridlock: Unraveling Explainability Bottlenecks in Automated Driving". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 7346–7355 (cit. on p. 84).

[74] DM Eddy, W Hollingworth, JJ Caro, et al. "Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7". In: *Value Health* 15.6 (2012), pp. 843–850 (cit. on p. 18).

[75] Burak Ekim, Timo T Stomberg, Ribana Roscher, and Michael Schmitt. "MapInWild: A remote sensing dataset to address the question of what makes nature wild [Software and Data Sets]". In: *IEEE Geoscience and Remote Sensing Magazine* 11.1 (2023), pp. 103–114 (cit. on p. 20).

[76] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. "ILIME: local and global interpretable model-agnostic explainer of black-box decision". In: *Advances in Databases and Information Systems: 23rd European Conference, ADBIS 2019, Bled, Slovenia, September 8–11, 2019, Proceedings 23*. Springer. 2019, pp. 53–68 (cit. on p. 19).

[77] TA Essa. "Effect of salinity stress on growth and nutrient composition of three soybean (Glycine max L. Merrill) cultivars". In: *Journal of Agronomy and Crop science* 188.2 (2002), pp. 86–93 (cit. on p. 149).

[78] John E Estes, Charlene Sailer, and Larry R Tinney. "Applications of artificial intelligence techniques to remote sensing". In: *The Professional Geographer* 38.2 (1986), pp. 133–141 (cit. on p. 13).

[79] Sophia Falk and Aimee van Wynsberghe. "Challenging AI for Sustainability: what ought it mean?" In: *AI and Ethics* (2023), pp. 1–11 (cit. on p. 106).

[80] Runyu Fan, Jun Li, Weijing Song, et al. "Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data". In: *International Journal of Applied Earth Observation and Geoinformation* 111 (2022), p. 102831 (cit. on p. 15).

[81]Nizom Farmonov, Khilola Amankulova, József Szatmári, et al. "Combining PlanetScope and Sentinel-2 Images with Environmental Data for Improved Wheat Yield Estimation". In: *International Journal of Digital Earth* 16.1 (2023), pp. 847–867 (cit. on p. 119).

[82]Tom G Farr and Mike Kobrick. "Shuttle Radar Topography Mission produces a wealth of data". In: *Eos, Transactions American Geophysical Union* 81.48 (2000), pp. 583–585 (cit. on p. 36).

[83]Z Fazakas, M Nilsson, and H Olsson. "Regional forest biomass and wood volume estimation using satellite data and ancillary data". In: *Agricultural and forest meteorology* 98 (1999), pp. 417–425 (cit. on p. 13).

[84]Felipe Ferrari, Matheus Pinheiro Ferreira, Cláudio Aparecido Almeida, and Raul Queiroz Feitosa. "Fusing Sentinel-1 and Sentinel-2 images for deforestation detection in the Brazilian amazon under diverse cloud conditions". In: *IEEE Geoscience and Remote Sensing Letters* 20 (2023), pp. 1–5 (cit. on p. 16).

[85]Gunther Fischer, FO Nachtergaele, HT Van Velthuizen, et al. *Global agro-ecological zones v4-model documentation*. Food & Agriculture Org., 2021 (cit. on p. 34).

[86]Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously". In: *Journal of Machine Learning Research* 20.177 (2019), pp. 1–81 (cit. on p. 21).

[87]Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3429–3437 (cit. on p. 21).

[88]Alex A Freitas. "Comprehensible classification models: a position paper". In: *ACM SIGKDD explorations newsletter* 15.1 (2014), pp. 1–10 (cit. on p. 4).

[89]MA Friedl, DS Schimel, J Michaelsen, FW Davis, and H Walker. "Estimating grassland biomass and leaf area index using ground and satellite data". In: *International Journal of Remote Sensing* 15.7 (1994), pp. 1401–1420 (cit. on p. 13).

[90]Mark A Friedl and Carla E Brodley. "Decision tree classification of land cover from remotely sensed data". In: *Remote sensing of environment* 61.3 (1997), pp. 399–409 (cit. on p. 13).

[91]Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 21).

[92]KS Fu, DA Landgrebe, and TL Phillips. "Information processing of remotely sensed agricultural data". In: *Proceedings of the IEEE* 57.4 (1969), pp. 639–653 (cit. on p. 13).

[93]Dinaol Gadisa and Hyochoong Bang. "Small satellite electro-optical system (EOS) technological and commercial expansion". In: *Acta Astronautica* (2023) (cit. on p. 11).

[94]JN Gallagher and PV Biscoe. "A physiological analysis of cereal yield. II. Partitioning of dry matter". In: *Agricultural progress* 53 (1978), pp. 51–70 (cit. on p. 49).

[95]V. Sainte Fare Garnot, L. Landrieu, S. Giordano, and N. Chehata. "Time-Space Tradeoff in Deep Learning Models for Crop Classification on Satellite Multi-Spectral Image Time Series". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 6247–6250 (cit. on p. 121).

[96] Vivien Sainte Fare Garnot and Loic Landrieu. "Lightweight temporal self-attention for classifying satellite images time series". In: *Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6*. Springer. 2020, pp. 171–181 (cit. on p. 109).

[97] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. "Multi-modal temporal attention models for crop mapping from satellite time series". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 187 (2022), pp. 294–305 (cit. on p. 16).

[98] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. "Satellite image time series classification with pixel-set encoders and temporal self-attention". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12325–12334 (cit. on pp. 57, 109).

[99] Mohamed Sadok Gastli, Lobna Nassar, and Fakhri Karray. "Satellite images and deep learning tools for crop yield prediction and price forecasting". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8 (cit. on p. 14).

[100] Jakob Gawlikowski, Patrick Ebel, Michael Schmitt, and Xiao Xiang Zhu. "Explaining the effects of clouds on remote sensing scene classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 9976–9986 (cit. on p. 23).

[101] Scott Gay, DB Egli, and DA Reicosky. "Physiological aspects of yield improvement in soybeans 1". In: *Agronomy Journal* 72.2 (1980), pp. 387–391 (cit. on p. 49).

[102] Ji Ge, Hong Zhang, Lu Xu, Chun-Ling Sun, and Chao Wang. "Interpretable Deep Learning Method Combining Temporal Backscattering Coefficients and Interferometric Coherence for Rice Area Mapping". In: *IEEE Geoscience and Remote Sensing Letters* (2023) (cit. on p. 24).

[103] Gohar Ghazaryan, Sergii Skakun, Simon König, et al. "Crop yield estimation using multi-source satellite image series and deep learning". In: *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2020, pp. 5163–5166 (cit. on p. 15).

[104] Bardan Ghimire, John Rogan, Víctor Rodríguez Galiano, Prajjwal Panday, and Neeti Neeti. "An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA". In: *GI-Science & Remote Sensing* 49.5 (2012), pp. 623–643 (cit. on p. 13).

[105] Amirata Ghorbani, Abubakar Abid, and James Zou. "Interpretation of neural networks is fragile". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 3681–3688 (cit. on pp. 22, 71).

[106] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, et al. "Contextual inter-modal attention for multi-modal sentiment analysis". In: *proceedings of the 2018 conference on empirical methods in natural language processing*. 2018, pp. 3454–3466 (cit. on p. 57).

[107] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. "Random forests for land cover classification". In: *Pattern recognition letters* 27.4 (2006), pp. 294–300 (cit. on p. 13).

[108] Anatoly Gitelson and Mark N Merzlyak. "Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves". In: *Journal of Photochemistry and Photobiology B: Biology* 22.3 (1994), pp. 247–252 (cit. on p. 124).

[109] Anatoly A Gitelson. "Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation". In: *Journal of plant physiology* 161.2 (2004), pp. 165–173 (cit. on p. 142).

[110] Anatoly A Gitelson, Yuri Gritz, and Mark N Merzlyak. "Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves". In: *Journal of plant physiology* 160.3 (2003), pp. 271–282 (cit. on p. 142).

[111] Anatoly A Gitelson, Yoram J Kaufman, and Mark N Merzlyak. "Use of a green channel in remote sensing of global vegetation from EOS-MODIS". In: *Remote sensing of Environment* 58.3 (1996), pp. 289–298 (cit. on p. 142).

[112] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation". In: *journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65 (cit. on p. 21).

[113] Gene H Golub and John H Welsch. "Calculation of Gauss quadrature rules". In: *Mathematics of computation* 23.106 (1969), pp. 221–230 (cit. on p. 31).

[114] David G Goodenough, Morris Goldberg, Gordon Plunkett, and John Zelek. "An expert system for remote sensing". In: *IEEE Transactions on Geoscience and Remote Sensing* 3 (1987), pp. 349–359 (cit. on p. 13).

[115] Riccardo Guidotti. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery* 38.5 (2024), pp. 2770–2824 (cit. on p. 19).

[116] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42 (cit. on pp. 4, 40).

[117] Alexander Günther, Hiba Najjar, and Andreas Dengel. "Explainable Multi-Modal Learning in Remote Sensing: Challenges and Future Directions". In: *IEEE Geoscience and Remote Sensing Letters* (2024) (cit. on p. 83).

[118] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines". In: *Machine learning* 46 (2002), pp. 389–422 (cit. on p. 106).

[119] Amir Haghverdi, Robert A Washington-Allen, and Brian G Leib. "Prediction of cotton lint yield from phenology of crop indices using artificial neural networks". In: *Computers and Electronics in Agriculture* 152 (2018), pp. 186–197 (cit. on p. 14).

[120] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. "Trusted multi-view classification with dynamic evidential fusion". In: *IEEE transactions on pattern analysis and machine intelligence* 45.2 (2022), pp. 2551–2566 (cit. on p. 16).

[121] Matthew Hansen, R Dubayah, and R DeFries. "Classification trees: an alternative to traditional land cover classifiers". In: *International journal of remote sensing* 17.5 (1996), pp. 1075–1081 (cit. on p. 13).

[122] Mahya GZ Hashemi, Pang-Ning Tan, Ehsan Jalilvand, et al. "Yield estimation from SAR data using patch-based deep learning and machine learning techniques". In: *Computers and Electronics in Agriculture* 226 (2024), p. 109340 (cit. on p. 14).

[123] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, et al. "Interpreting black-box models: a review on explainable artificial intelligence". In: *Cognitive Computation* 16.1 (2024), pp. 45–74 (cit. on pp. 17, 19).

[124] Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*. 2009 (cit. on p. 21).

[125] Patrick Helber, Benjamin Bischke, Peter Habelitz, et al. "Crop Yield Prediction: An Operational Approach to Crop Yield Modeling on Field and Subfield Level with Machine Learning Models". In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. 2023 (cit. on p. 58).

[126] Patrick Helber, Benjamin Bischke, Carolin Packbier, Peter Habelitz, and Florian Seefeldt. "An Operational Approach to Large-Scale Crop Yield Prediction with Spatio-Temporal Machine Learning Models". In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2024, pp. 4299–4302 (cit. on p. 58).

[127] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, et al. "Generating visual explanations". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 3–19 (cit. on p. 84).

[128] Georgef Hepner, Thomas Logan, Niles Ritter, and Nevin Bryant. "Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification". In: *Photogrammetric Engineering and Remote Sensing* 56.4 (1990), pp. 469–473 (cit. on p. 13).

[129] Hans Hersbach, Bill Bell, Paul Berrisford, et al. "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730 (2020), pp. 1999–2049 (cit. on p. 36).

[130] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models". In: *Advances in neural information processing systems* 33 (2020), pp. 4778–4789 (cit. on p. 21).

[131] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on pp. 38, 60, 109).

[132] Adrian Höhl, Ivica Obadic, Miguel-Ángel Fernández-Torres, et al. "Opening the Black Box: A systematic review on explainable artificial intelligence in remote sensing". In: *IEEE Geoscience and Remote Sensing Magazine* (2024) (cit. on pp. 21, 23, 25).

[133] Danfeng Hong, Jocelyn Chanussot, Naoto Yokoya, et al. "WU-Net: A weakly-supervised unmixing network for remotely sensed hyperspectral imagery". In: *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*. IEEE. 2019, pp. 373–376 (cit. on p. 15).

[134] Danfeng Hong, Lianru Gao, Naoto Yokoya, et al. "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.5 (2020), pp. 4340–4354 (cit. on p. 16).

[135] Danfeng Hong, Naoto Yokoya, Jocelyn Chanussot, Jian Xu, and Xiao Xiang Zhu. "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction". In: *ISPRS journal of photogrammetry and remote sensing* 158 (2019), pp. 35–49 (cit. on p. 15).

[136] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. "A benchmark for interpretability methods in deep neural networks". In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 106, 107, 110, 114).

[137]Bo Huang, Bei Zhao, and Yimeng Song. "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery". In: *Remote Sensing of Environment* 214 (2018), pp. 73–86 (cit. on p. 15).

[138]Yu Huang, Chenzhuang Du, Zihui Xue, et al. "What makes multi-modal learning better than single (provably)". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10944–10956 (cit. on p. 83).

[139]Alfredo Huete, Kamel Didan, Tomoaki Miura, et al. "Overview of the radiometric and biophysical performance of the MODIS vegetation indices". In: *Remote sensing of environment* 83.1-2 (2002), pp. 195–213 (cit. on p. 119).

[140]Alfredo R Huete. "A soil-adjusted vegetation index (SAVI)". In: *Remote sensing of environment* 25.3 (1988), pp. 295–309 (cit. on p. 119).

[141]Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models". In: *Decision Support Systems* 51.1 (2011), pp. 141–154 (cit. on p. 4).

[142]Alvin Inderka, Florian Huber, and Volker Steinhage. "On Convolutional Vision Transformers for Yield Prediction". In: *arXiv preprint arXiv:2402.05557* (2024) (cit. on p. 58).

[143]Yoshio Inoue. "Satellite-and drone-based remote sensing of crops and soils for smart farming–a review". In: *Soil Science and Plant Nutrition* 66.6 (2020), pp. 798–810 (cit. on p. 12).

[144]R. Interdonato, D. Ienco, R. Gaetano, and K. Ose. "DuPLO: A DUal view Point deep Learning architecture for time series classificatiOn". In: *ISPRS journal of photogrammetry and remote sensing* 149 (2019), pp. 91–104 (cit. on p. 121).

[145]Shin-nosuke Ishikawa, Masato Todo, Masato Taki, et al. "Example-based explainable AI and its application for remote sensing image classification". In: *International Journal of Applied Earth Observation and Geoinformation* 118 (2023), p. 103215 (cit. on p. 20).

[146]Alon Jacovi and Yoav Goldberg. "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" In: *arXiv preprint arXiv:2004.03685* (2020) (cit. on p. 25).

[147]Giyoung Jeon, Haedong Jeong, and Jaesik Choi. "Distilled gradient aggregation: Purify features for input attribution in the deep neural network". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26478–26491 (cit. on p. 19).

[148]Jennifer L Jewiss, Molly E Brown, and Vanessa M Escobar. "Satellite remote Sensing data for decision Support in emerging Agricultural economies: How Satellite data can transform Agricultural decision making [Perspectives]". In: *IEEE Geoscience and Remote Sensing Magazine* 8.4 (2020), pp. 117–133 (cit. on p. 12).

[149]Sen Jia, Zhihao Wang, Qingquan Li, Xiuping Jia, and Meng Xu. "Multiattention generative adversarial network for remote sensing image superresolution". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–15 (cit. on p. 15).

[150]X. Jia, A. Khandelwal, G. Nayak, et al. "Incremental Dual-memory LSTM in Land Cover Prediction". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 867–876 (cit. on p. 121).

[151] Hao Jiang, Hao Hu, Renhai Zhong, et al. "A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level". In: *Global change biology* 26.3 (2020), pp. 1754–1766 (cit. on p. 45).

[152] Li Jianguo and Mao Jietai. "The approach to remote sensing of water vapor based on GPS and linear regression Tm in eastern region of China". In: *Journal of Meteorological Research* 12.4 (1998), pp. 450–458 (cit. on p. 13).

[153] Gong Jianya, Sui Haigang, Ma Guorui, and Zhou Qiming. "A review of multi-temporal remote sensing data change detection algorithms". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 37.B7 (2008), pp. 757–762 (cit. on p. 13).

[154] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. "A review on explainability in multimodal deep neural nets". In: *IEEE Access* 9 (2021), pp. 59800–59821 (cit. on pp. 57, 83).

[155] Tejas Junankar, Jasleen Kaur Sondhi, and Akhil M Nair. "Wheat Yield Prediction using Temporal Fusion Transformers". In: *2023 2nd International Conference for Innovation in Technology (INOCON)*. IEEE. 2023, pp. 1–6 (cit. on p. 58).

[156] Heechul Jung and Taegyun Jeon. "Self-supervised learning with randomised layers for remote sensing". In: *Electronics Letters* 57.6 (2021), pp. 249–251 (cit. on p. 15).

[157] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 1–9 (cit. on p. 15).

[158] Atsushi Kanehira and Tatsuya Harada. "Learning to explain with complemental examples". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 8603–8611 (cit. on p. 84).

[159] B Kartikeyan, Kantilal L Majumder, and AR Dasgupta. "An expert system for land cover classification". In: *IEEE Transactions on geoscience and remote sensing* 33.1 (1995), pp. 58–66 (cit. on p. 13).

[160] Alex Kendall, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491 (cit. on pp. 83, 102).

[161] Patrick W Keys, Elizabeth A Barnes, and Neil H Carter. "A machine-learning approach to human footprint index estimation with applications to sustainable development". In: *Environmental Research Letters* 16.4 (2021), p. 044061 (cit. on p. 20).

[162] Mehak Khan, Abdul Hanan, Meruyert Kenzhebay, Michele Gazzea, and Reza Arghandeh. "Transformer-based land use and land cover classification with explainability using satellite imagery". In: *Scientific Reports* 14.1 (2024), p. 16744 (cit. on p. 57).

[163] Taewoo Kim, Minsu Jeon, Changha Lee, et al. "Federated onboard-ground station computing with weakly supervised cascading pyramid attention network for satellite image analysis". In: *IEEE Access* 10 (2022), pp. 117315–117333 (cit. on pp. 20, 57).

[164]Nobuyuki Kobayashi, Hiroshi Tani, Xiufeng Wang, and Rei Sonobe. "Crop classification using spectral indices derived from Sentinel-2A imagery". In: *Journal of Information and Telecommunication* 4.1 (2020), pp. 67–90 (cit. on p. 120).

[165]Narine Kokhlikyan, Vivek Miglani, Miguel Martin, et al. "Captum: A unified and generic model interpretability library for pytorch". In: *arXiv preprint arXiv:2009.07896* (2020) (cit. on p. 31).

[166]Yingying Kong, Siyuan Liu, and Xiangyang Peng. "Multi-Scale translation method from SAR to optical remote sensing images based on conditional generative adversarial network". In: *International Journal of Remote Sensing* 43.8 (2022), pp. 2837–2860 (cit. on p. 15).

[167]Herbert J Kramer et al. *Observation of the Earth and its Environment: Survey of Missions and Sensors*. Vol. 1982. Springer, 2002 (cit. on p. 11).

[168]V Gokula Krishnan, BV Subba Rao, J Rajendra Prasad, P Pushpa, and S Kumari. "Sugarcane yield prediction using NOA-based swin transformer model in IoT smart agriculture". In: *Journal of Applied Biology and Biotechnology* 12.2 (2024), pp. 239–247 (cit. on p. 58).

[169]Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012) (cit. on p. 15).

[170]Saatha Kumudini. "Soybean growth and development." In: *The soybean: botany, production and uses*. CABI Wallingford UK, 2010, pp. 48–73 (cit. on pp. 48, 49, 147, 149).

[171]Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. "Deep learning classification of land cover and crop types using remote sensing data". In: *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017), pp. 778–782 (cit. on p. 15).

[172]Heesung Kwon and Nasser M Nasrabadi. "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery". In: *IEEE transactions on Geoscience and Remote Sensing* 43.2 (2005), pp. 388–397 (cit. on p. 13).

[173]Peter D Lancashire, Hermann Bleiholder, T Van Den Boom, et al. "A uniform decimal code for growth stages of crops and weeds". In: *Annals of applied Biology* 119.3 (1991), pp. 561–601 (cit. on pp. 41, 141).

[174]David Landgrebe. "Computer-based remote sensing technology-a look to the future". In: *Remote Sensing of Environment* 5 (1976), pp. 229–246 (cit. on p. 13).

[175]David A Landgrebe. "Automatic identification and classification of wheat by remote sensing". In: (1967) (cit. on p. 13).

[176]Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. "Analyzing classifiers: Fisher vectors and deep neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2912–2920 (cit. on p. 22).

[177]Rick L Lawrence, Shana D Wood, and Roger L Sheley. "Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest)". In: *Remote Sensing of Environment* 100.3 (2006), pp. 356–362 (cit. on p. 13).

[178]Yann LeCun, Bernhard Boser, John S Denker, et al. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551 (cit. on p. 15).

[179] Alex Levering, Diego Marcos, and Devis Tuia. "On the relation between landscape beauty and land cover: A case study in the UK at Sentinel-2 resolution with interpretable AI". In: *ISPRS journal of Photogrammetry and Remote Sensing* 177 (2021), pp. 194–203 (cit. on pp. 83, 84, 102).

[180] Elissa R Levine, Lubomir Kurz, Jan Smid, Marek Smid, and Petr Volf. "Algorithms and analysis tools for carbon content modeling in soil based on satellite data". In: *Remote Sensing for Agriculture, Ecosystems, and Hydrology*. Vol. 3499. SPIE. 1998, pp. 315–322 (cit. on p. 13).

[181] Jiwei Li, Will Monroe, and Dan Jurafsky. "Understanding neural networks through representation erasure". In: *arXiv preprint arXiv:1612.08220* (2016) (cit. on p. 21).

[182] Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. "Geographical knowledge-driven representation learning for remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), pp. 1–16 (cit. on p. 15).

[183] Xinghua Li, Zhengshun Du, Yanyuan Huang, and Zhenyu Tan. "A deep translation (GAN) based change detection network for optical and SAR remote sensing images". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 179 (2021), pp. 14–34 (cit. on p. 15).

[184] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. "A survey of convolutional neural networks: analysis, applications, and prospects". In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019 (cit. on p. 15).

[185] Zhicheng Li and Laurent Itti. "Saliency and gist features for target detection in satellite images". In: *IEEE Transactions on Image Processing* 20.7 (2010), pp. 2017–2029 (cit. on p. 13).

[186] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. "Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions". In: *ACM Comput. Surv.* 56.10 (2024) (cit. on p. 62).

[187] Fudong Lin, Summer Crawford, Kaleb Guillot, et al. "MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 5774–5784 (cit. on p. 58).

[188] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57 (cit. on p. 18).

[189] Hui Liu, Qingyu Yin, and William Yang Wang. "Towards Explainable NLP: A Generative Explanation Framework for Text Classification". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 5570–5581 (cit. on p. 84).

[190] Shikun Liu, Edward Johns, and Andrew J Davison. "End-to-end multi-task learning with attention". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1871–1880 (cit. on pp. 83, 102).

[191] Y. Liu, J. Qian, and H. Yue. "Comprehensive evaluation of Sentinel-2 red edge and shortwave-infrared bands to estimate soil moisture". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), pp. 7448–7465 (cit. on p. 122).

[192]Yu Liu and Tinne Tuytelaars. "A deep multi-modal explanation model for zero-shot learning". In: *IEEE Transactions on Image Processing* 29 (2020), pp. 4788–4803 (cit. on p. 84).

[193]CP Lo. "Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach". In: *Remote Sensing* 16.1 (1995), pp. 17–34 (cit. on p. 13).

[194]Max Losch, Mario Fritz, and Bernt Schiele. "Interpretability beyond classification output: Semantic bottleneck networks". In: *arXiv preprint arXiv:1907.10882* (2019) (cit. on p. 84).

[195]I Loshchilov. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017) (cit. on p. 63).

[196]Ilya Loshchilov and Frank Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts". In: *International Conference on Learning Representations*. 2022 (cit. on p. 63).

[197]Dengsheng Lu, Paul Mausel, Eduardo Brondizio, and Emilio Moran. "Change detection techniques". In: *International journal of remote sensing* 25.12 (2004), pp. 2365–2401 (cit. on p. 13).

[198]Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. "12-in-1: Multi-task vision and language representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10437–10446 (cit. on pp. 83, 102).

[199]Scott M Lundberg, Gabriel G Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles". In: *arXiv preprint arXiv:1802.03888* (2018) (cit. on pp. 21, 43).

[200]Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 21, 22, 31, 32, 57, 121).

[201]Yuchi Ma, Zhou Zhang, Yanghui Kang, and Mutlu Özdoğan. "Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach". In: *Remote Sensing of Environment* 259 (2021), p. 112408 (cit. on p. 45).

[202]Donato Malerba and Vincenzo Pasquadibisceglie. "Data-Centric AI". In: *Journal of Intelligent Information Systems* (2024), pp. 1–10 (cit. on pp. 19, 105).

[203]Austen Maniscalco, Ezek Mathew, David Parsons, et al. "Multimodal radiotherapy dose prediction using a multi-task deep learning model". In: *Medical physics* (2024) (cit. on pp. 83, 102).

[204]Diego Marcos, Sylvain Lobry, and Devis Tuia. "Semantically Interpretable Activation Maps: what-where-how explanations within CNNs". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 4207–4215 (cit. on p. 84).

[205]Mohammad Marjani, Fariba Mohammadimanesh, Masoud Mahdianpari, and Eric W Gill. "A Novel Spatio-Temporal Vision Transformer Model for Improving Wetland Mapping Using Multi-Seasonal Sentinel Data". In: *Remote Sensing Applications: Society and Environment* (2024), p. 101401 (cit. on p. 15).

[206]Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. "Deep learning earth observation classification using ImageNet pretrained networks". In: *IEEE Geoscience and Remote Sensing Letters* 13.1 (2015), pp. 105–109 (cit. on p. 15).

[207] Ali Masjedi, Neal R Carpenter, Melba M Crawford, and Mitch R Tuinstra. "Prediction of sorghum biomass using UAV time series data and recurrent neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0 (cit. on p. 15).

[208] Heather McNairn, Angela Kross, David Lapen, Ron Caves, and Jiali Shang. "Early season monitoring of corn and soybeans with TerraSAR-X and RADARSAT-2". In: *International Journal of Applied Earth Observation and Geoinformation* 28 (2014), pp. 252–259 (cit. on p. 14).

[209] Denise A McWilliams, Duane Raymond Berglund, and GJ Endres. *Soybean growth and management quick guide*. North Dakota State University, 1999 (cit. on pp. 41, 49, 141, 148, 149).

[210] N Méger, H Courteille, A Benoit, A Atto, and Dino Ienco. "Explaining a deep spatiotemporal land cover classifier with attention and redescription mining". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43 (2022), pp. 673–680 (cit. on p. 57).

[211] Francisco Mena, Diego Arenas, Marlon Nuske, and Andreas Dengel. "Common Practices and Taxonomy in Deep Multi-view Fusion for Remote Sensing Applications". In: *arXiv preprint arXiv:2301.01200* (2023) (cit. on p. 39).

[212] Francisco Mena, Diego Arenas, Marlon Nuske, and Andreas Dengel. "Common practices and taxonomy in deep multi-view fusion for remote sensing applications". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024) (cit. on pp. 15, 16, 62, 88).

[213] Francisco Mena, Deepak Pathak, Hiba Najjar, et al. "Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction". In: *Remote Sensing of Environment* 318 (2025), p. 114547 (cit. on pp. 39, 58, 80).

[214] Joel Michaelsen, David S Schimel, Mark A Friedl, Frank W Davis, and Ralph C Dubayah. "Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys". In: *Journal of Vegetation Science* 5.5 (1994), pp. 673–686 (cit. on p. 13).

[215] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38 (cit. on p. 16).

[216] LK Milne, TD Gedeon, and AK Skidmore. "Classifying Dry Sclerophyll Forest from Augmented Satellite Data: Comparing Neural Network, Decision Tree & Maximum likelihood". In: *Proceedings of the Australian Conference on Neural Networks*. 1995, pp. 160–163 (cit. on p. 13).

[217] Miro Miranda, Deepak Pathak, Marlon Nuske, and Andreas Dengel. "Multi-Modal Fusion Methods with Local Neighborhood Information for Crop Yield Prediction at Field and Subfield Levels". In: *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2024, pp. 4307–4311 (cit. on p. 39).

[218] G. Misra, F. Cawkwell, and A. Wingler. "Status of phenological research using Sentinel-2 data: A review". In: *Remote Sensing* 12.17 (2020), p. 2760 (cit. on pp. 119, 126).

[219] Nooshin Mojab, Vahid Noroozi, S Yu Philip, and Joelle A Hallak. "Deep multi-task learning for interpretable glaucoma detection". In: *2019 IEEE 20th International conference on information reuse and integration for data science (IRI)*. IEEE. 2019, pp. 167–174 (cit. on p. 84).

[220]Michael Mommert, Nicolas Kesseli, Joëlle Hanna, et al. "Ben-ge: Extending BigEarthNet with geographical and environmental data". In: *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2023, pp. 1016–1019 (cit. on pp. 86, 91).

[221]Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. "Layer-wise relevance propagation: an overview". In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 193–209 (cit. on p. 22).

[222]Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks". In: *Digital signal processing* 73 (2018), pp. 1–15 (cit. on pp. 18, 19).

[223]Aaron Moody and Curtis E Woodcock. "The influence of scale and the spatial characteristics of landscapes on land-cover mapping using remote sensing". In: *Landscape Ecology* 10 (1995), pp. 363–379 (cit. on p. 13).

[224]Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 607–617 (cit. on p. 19).

[225]L. Mou, L. Bruzzone, and X. X. Zhu. "Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.2 (2018), pp. 924–935 (cit. on p. 121).

[226]L. Mou, P. Ghamisi, and X. X. Zhu. "Deep Recurrent Neural Networks for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.7 (2017), pp. 3639–3655 (cit. on p. 121).

[227]W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. "Interpretable machine learning: definitions, methods, and applications". In: *arXiv preprint arXiv:1901.04592* (2019) (cit. on pp. 18, 19).

[228]Priyanga Muruganantham, Santoso Wibowo, Srimannarayana Grandhi, Nahidul Hoque Samrat, and Nahina Islam. "A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing". In: *Remote Sensing* 14.9 (2022), p. 1990 (cit. on p. 45).

[229]Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 807–814 (cit. on p. 38).

[230]Hiba Najjar, Marlon Nuske, and Andreas Dengel. "Data-Centric Machine Learning for Earth Observation: Necessary and Sufficient Features". In: *arXiv preprint arXiv:2408.11384* (2024) (cit. on p. 50).

[231]Marla Narazani, Ignacio Sarasua, Sebastian Pölsterl, et al. "Is a pet all you need? a multi-modal study for alzheimer's disease using 3d cnns". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 66–76 (cit. on p. 83).

[232]Meike Nauta, Jan Trienes, Shreyasi Pathak, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai". In: *ACM Computing Surveys* 55.13s (2023), pp. 1–42 (cit. on pp. 25, 106).

[233]Petteri Nevavuori, Nathaniel Narra, Petri Linna, and Tarmo Lipping. "Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models". In: *Remote sensing* 12.23 (2020), p. 4000 (cit. on p. 45).

[234] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26422–26436 (cit. on pp. 19, 23).

[235] Keiller Nogueira, Mauro Dalla Mura, Jocelyn Chanussot, William Robson Schwartz, and Jefersson Alex Dos Santos. "Dynamic multicontext segmentation of remote sensing images based on convolutional networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.10 (2019), pp. 7503–7520 (cit. on p. 15).

[236] OpenStreetMap contributors. *OpenStreetMap*. Accessed: 2025-02-17. 2025 (cit. on pp. 35, 60).

[237] Aiym Orynbaikyzy, Ursula Gessner, and Christopher Conrad. "Crop type classification using a combination of optical and radar remote sensing data: A review". In: *international journal of remote sensing* 40.17 (2019), pp. 6553–6595 (cit. on p. 12).

[238] Amir Hosein Oveis, Elisa Giusti, Selenia Ghio, Giulio Meucci, and Marco Martorella. "LIME-Assisted Automatic Target Recognition with SAR Images: Towards Incremental Learning and Explainability". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2023) (cit. on p. 24).

[239] Mahesh Pal. "Random forest classifier for remote sensing classification". In: *International journal of remote sensing* 26.1 (2005), pp. 217–222 (cit. on p. 13).

[240] Sebastian Palacio, Adriano Lucieri, Mohsin Munir, et al. "Xai handbook: towards a unified framework for explainable AI". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3766–3775 (cit. on p. 17).

[241] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, et al. "Multimodal explanations: Justifying decisions and pointing to the evidence". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8779–8788 (cit. on p. 84).

[242] Hugh Pasika, Simon Haykin, Eugene Clothiaux, and Ron Stewart. "Neural networks for sensor fusion in remote sensing". In: *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*. Vol. 4. IEEE. 1999, pp. 2772–2776 (cit. on p. 13).

[243] Deepak Pathak, Miro Miranda, Francisco Mena, et al. "Predicting Crop Yield With Machine Learning: An Extensive Analysis Of Input Modalities And Models On a Field and Subfield Level". In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. 2023 (cit. on pp. 36, 39).

[244] Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. "Temporal convolutional neural network for the classification of satellite image time series". In: *Remote Sensing* 11.5 (2019), p. 523 (cit. on p. 61).

[245] Otávio AB Penatti, Keiller Nogueira, and Jefersson A Dos Santos. "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 44–51 (cit. on p. 15).

[246] Gregor Perich, Mehmet Ozgur Turkoglu, Lukas Valentin Graf, et al. "Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks". In: *Field Crops Research* 292 (2023), p. 108824 (cit. on p. 108).

[247]Ph Puyou-Lascassies, G Flouzat, M Gay, and C Vignolles. "Validation of the use of multiple linear regression as a tool for unmixing coarse spatial resolution images". In: *Remote Sensing of environment* 49.2 (1994), pp. 155–166 (cit. on p. 13).

[248]JongCheol Pyo, Kyung Hwa Cho, Kyunghyun Kim, et al. "Cyanobacteria cell prediction using interpretable deep learning model with observed, numerical, and sensing data assemblage". In: *Water Research* 203 (2021), p. 117483 (cit. on p. 57).

[249]Mengjia Qiao, Xiaohui He, Xijie Cheng, et al. "KSTAGE: A knowledge-guided spatial-temporal attention graph learning network for crop yield prediction". In: *Information Sciences* 619 (2023), pp. 19–37 (cit. on p. 58).

[250]Balaji Rajagopalan and Upmanu Lall. "A k-nearest-neighbor simulator for daily precipitation and other weather variables". In: *Water resources research* 35.10 (1999), pp. 3089–3101 (cit. on p. 13).

[251]Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. "Explainable deep learning: A field guide for the uninitiated". In: *Journal of Artificial Intelligence Research* 73 (2022), pp. 329–396 (cit. on pp. 18, 19, 21, 23, 32).

[252]Bruce H Raup, Liss M Andreassen, Tobias Bolch, and Suzanne Bevan. "Remote sensing of glaciers". In: *Remote Sensing of the Cryosphere* (2015), pp. 123–156 (cit. on p. 12).

[253]Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144 (cit. on pp. 22, 32).

[254]Isabel Rio-Torto, Kelwin Fernandes, and Luís F Teixeira. "Understanding the decisions of CNNs: An in-model approach". In: *Pattern Recognition Letters* 133 (2020), pp. 373–380 (cit. on p. 84).

[255]Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. "Explainable machine learning for scientific insights and discoveries". In: *Ieee Access* 8 (2020), pp. 42200–42216 (cit. on pp. 18, 19).

[256]Ribana Roscher, Marc Rußwurm, Caroline Gevaert, et al. "Better, not just more: Data-centric machine learning for Earth observation". In: *IEEE Geoscience and Remote Sensing Magazine* (2024) (cit. on pp. 19, 105).

[257]John Wilson Rouse, Rüdiger H Haas, John A Schell, Donald W Deering, et al. "Monitoring vegetation systems in the Great Plains with ERTS". In: *NASA Spec. Publ* 351.1 (1974), p. 309 (cit. on pp. 119, 124, 142).

[258]Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215 (cit. on pp. 17, 82).

[259]Marc Rußwurm and Marco Körner. "Self-attention for raw optical satellite time series classification". In: *ISPRS journal of photogrammetry and remote sensing* 169 (2020), pp. 421–435 (cit. on pp. 20, 57, 68).

[260]R. Rustowicz, R. Cheong, L. Wang, et al. "Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 75–82 (cit. on p. 120).

[261]Maria Sahakyan, Zeyar Aung, and Talal Rahwan. "Explainable artificial intelligence for tabular data: A survey". In: *IEEE access* 9 (2021), pp. 135392–135422 (cit. on p. 24).

[262] Arnt-Børre Salberg. "Detection of seals in remote sensing images using features extracted from deep convolutional neural networks". In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2015, pp. 1893–1896 (cit. on p. 15).

[263] Tobias Sauter, Björn Weitzenkamp, and Christoph Schneider. "Spatio-temporal prediction of snow cover in the Black Forest mountain range using remote sensing and a recurrent neural network". In: *International Journal of Climatology* 30.15 (2010), pp. 2330–2341 (cit. on p. 15).

[264] A Schaum. "Joint subspace detection of hyperspectral targets". In: *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No. 04TH8720)*. Vol. 3. IEEE. 2004 (cit. on p. 13).

[265] AJ Schweiger and JR Key. "Estimating surface radiation fluxes in the Arctic from TOVS brightness temperatures". In: *International Journal of Remote Sensing* 18.4 (1997), pp. 955–970 (cit. on p. 13).

[266] Grant J Scott, Matthew R England, William A Starms, Richard A Marcum, and Curt H Davis. "Training deep convolutional neural networks for land–cover classification of high-resolution imagery". In: *IEEE Geoscience and Remote Sensing Letters* 14.4 (2017), pp. 549–553 (cit. on p. 15).

[267] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626 (cit. on pp. 23, 79, 121).

[268] Ozan Sener and Vladlen Koltun. "Multi-task learning as multi-objective optimization". In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 83, 102).

[269] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on p. 69).

[270] Zhenfeng Shao, Yin Pan, Chunyuan Diao, and Jiajun Cai. "Cloud detection in remote sensing images based on multiscale features-convolutional neural network". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.6 (2019), pp. 4062–4076 (cit. on p. 15).

[271] LS SHAPLEY. "A value for n-person games". In: *Contributions to the Theory of Games* (1953), pp. 307–317 (cit. on p. 21).

[272] A. Sharma, X. Liu, and X. Yang. "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks". In: *Neural Networks* 105 (2018), pp. 346–355 (cit. on p. 121).

[273] Mengyun Shi, Fengying Xie, Yue Zi, and Jihao Yin. "Cloud detection of remote sensing images by deep learning". In: *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2016, pp. 701–704 (cit. on p. 15).

[274] Xingjian Shi, Zhourong Chen, Hao Wang, et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in neural information processing systems* 28 (2015) (cit. on p. 15).

[275] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. "Not just a black box: Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1605.01713* (2016) (cit. on pp. 22, 31).

[276]K Simonyan, A Vedaldi, and A Zisserman. "Deep inside convolutional networks: visualising image classification models and saliency maps". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR. 2014 (cit. on pp. 22, 31).

[277]Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017) (cit. on p. 109).

[278]James R Smith and Randall L Nelson. "Selection for Seed-Filling Period in Soybean 1". In: *Crop science* 26.3 (1986), pp. 466–469 (cit. on p. 49).

[279]Nguyen-Thanh Son, Chi-Farn Chen, Youg-Sin Cheng, et al. "Field-scale rice yield prediction from Sentinel-2 monthly image composites using machine learning algorithms". In: *Ecological informatics* 69 (2022), p. 101618 (cit. on p. 14).

[280]Rei Sonobe, Hiroshi Tani, Xiufeng Wang, Yasuhito Kojima, and Nobuyuki Kobayashi. "Extreme Learning Machine-based Crop Classification using ALOS/PALSAR Images". In: *Japan agricultural research quarterly: JARQ* 49.4 (2015), pp. 377–381 (cit. on p. 14).

[281]Timo Speith. "A review of taxonomies of explainable artificial intelligence (XAI) methods". In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2022, pp. 2239–2250 (cit. on pp. 16, 17).

[282]J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. "Striving for Simplicity: The All Convolutional Net". In: *ICLR (workshop track)*. 2015 (cit. on p. 109).

[283]Shivangi Srivastava, John E Vargas Munoz, Sylvain Lobry, and Devis Tuia. "Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data". In: *International Journal of Geographical Information Science* 34.6 (2020), pp. 1117–1136 (cit. on p. 16).

[284]Trevor Standley, Amir Zamir, Dawn Chen, et al. "Which tasks should be learned together in multi-task learning?" In: *International conference on machine learning*. PMLR. 2020, pp. 9120–9132 (cit. on p. 83).

[285]Sarah Sterz, Kevin Baum, Anne Lauber-Rönsberg, and Holger Hermanns. "Towards perspicuity requirements". In: *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE. 2021, pp. 159–163 (cit. on p. 16).

[286]Erik Strumbelj and Igor Kononenko. "An efficient explanation of individual classifications using game theory". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1–18 (cit. on pp. 21, 32, 109, 121).

[287]Martin Sudmanns, Dirk Tiede, Hannah Augustin, and Stefan Lang. "Assessing global Sentinel-2 coverage dynamics and data availability for operational Earth observation (EO) applications using the EO-Compass". In: *International journal of digital earth* 13.7 (2020), pp. 768–784 (cit. on p. 39).

[288]John J Sulik and Dan S Long. "Spectral considerations for modeling yield of canola". In: *Remote Sensing of Environment* 184 (2016), pp. 161–174 (cit. on p. 142).

[289]Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 5901–5904 (cit. on p. 86).

[290] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, et al. "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]". In: *IEEE Geoscience and Remote Sensing Magazine* 9.3 (2021), pp. 174–180 (cit. on p. 86).

[291] Weiwei Sun and Ruisheng Wang. "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM". In: *IEEE Geoscience and Remote Sensing Letters* 15.3 (2018), pp. 474–478 (cit. on p. 15).

[292] Xia Sun, Na Li, and Hui-jie Zhao. "Performance evaluation for hyperspectral target detection algorithms". In: *Seventh International Symposium on Instrumentation and Control Technology: Sensors and Instruments, Computer Simulation, and Artificial Intelligence*. Vol. 7127. SPIE. 2008, pp. 485–490 (cit. on p. 13).

[293] Mukund Sundararajan and Amir Najmi. "The many Shapley values for model explanation". In: *International conference on machine learning*. PMLR. 2020, pp. 9269–9278 (cit. on p. 21).

[294] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328 (cit. on pp. 22, 31, 57).

[295] Zheng Tang and Mihai Surdeanu. "It takes two flints to make a fire: Multitask learning of neural relation and explanation classifiers". In: *Computational Linguistics* 49.1 (2023), pp. 117–156 (cit. on p. 84).

[296] Maria Tattaris, Matthew P Reynolds, and Scott C Chapman. "A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding". In: *Frontiers in plant science* 7 (2016), p. 1131 (cit. on p. 14).

[297] Marco Tedesco. "Remote sensing and the cryosphere". In: *Remote Sensing of the Cryosphere* (2015), pp. 1–16 (cit. on p. 11).

[298] Marco Tedesco, Tommaso Parrinello, Charles Webb, and Thorsten Markus. "Remote sensing missions and the cryosphere". In: *Remote sensing of the cryosphere* (2015), pp. 382–392 (cit. on p. 12).

[299] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. "Shifting the Baseline: Single Modality Performance on Visual Navigation & QA". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 1977–1983 (cit. on p. 83).

[300] Allison M Thomson, Robert A Brown, Steven J Ghan, et al. "Elevation dependence of winter wheat production in eastern Washington State with climate change: A methodological study". In: *Climatic Change* 54 (2002), pp. 141–164 (cit. on p. 51).

[301] Huiren Tian, Pengxin Wang, Kevin Tansey, et al. "A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China". In: *International Journal of Applied Earth Observation and Geoinformation* 102 (2021) (cit. on pp. 58, 60).

[302] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288 (cit. on p. 106).

[303]R Tiinjes, S Glowe, J Biicknel, and C Lledtke. "Knowledge-based interpretation of remote sensing lmages using semantic nets". In: *Photogrammetric Engineering & Remote Sensing* 65.7 (1999), pp. 811–821 (cit. on p. 13).

[304]Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems". In: *arXiv preprint arXiv:1806.07552* (2018) (cit. on p. 18).

[305]Paul M Treitz, Philip J Howarth, and Peng Gong. "Application of satellite and GIS technologies for land-cover and land-use mapping at the rural-urban fringe: a case study". In: *Photogrammetric Engineering & Remote Sensing* (1992) (cit. on p. 13).

[306]Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, et al. "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for computational linguistics. Meeting*. Vol. 2019. NIH Public Access. 2019, p. 6558 (cit. on p. 57).

[307]M Tsekhmeistruk, O Pankova, V Kolomatska, et al. "Influence of weather and climatic conditions on soybean yield". In: *Ukrainian Journal of Ecology* 11.4 (2021), pp. 11–17 (cit. on p. 149).

[308]Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. "Cropharvest: A global dataset for crop-type classification". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021 (cit. on p. 108).

[309]Jingzhi Tu, Gang Mei, Zhengjing Ma, and Francesco Piccialli. "SWCGAN: Generative adversarial network combining swin transformer and CNN for remote sensing image super-resolution". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 5662–5673 (cit. on p. 15).

[310]Ziming Tu, Xiubin Yang, Xi He, Jiapu Yan, and Tingting Xu. "RGTGAN: Reference-Based Gradient-Assisted Texture-Enhancement GAN for Remote Sensing Super-Resolution". In: *IEEE Transactions on Geoscience and Remote Sensing* (2024) (cit. on p. 15).

[311]Compton J Tucker. "Red and photographic infrared linear combinations for monitoring vegetation". In: *Remote sensing of Environment* 8.2 (1979), pp. 127–150 (cit. on p. 142).

[312]Devis Tuia, Konrad Schindler, Begüm Demir, et al. "Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward". In: *IEEE Geoscience and Remote Sensing Magazine* (2024) (cit. on p. 3).

[313]Javier Noa Turnes, Jose David Bermudez Castro, Daliana Lobo Torres, et al. "Atrous cGAN for SAR to optical image translation". In: *IEEE Geoscience and Remote Sensing Letters* 19 (2020), pp. 1–5 (cit. on p. 15).

[314]Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, et al. "Multi-task learning for dense prediction tasks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3614–3633 (cit. on p. 83).

[315]Bhavan Vasu and Andreas Savakis. "Resilience and plasticity of deep network interpretations for aerial imagery". In: *IEEE Access* 8 (2020), pp. 127491–127506 (cit. on p. 23).

[316] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 60, 61, 68, 69).

[317] Emmanuelle Vaudour, Paul-Emile Noirot-Cosson, and Olivier Membrive. "Early-season mapping of crops and cultural operations using very high spatial resolution Pléiades images". In: *International Journal of Applied Earth Observation and Geoinformation* 42 (2015), pp. 128–141 (cit. on p. 14).

[318] Giulia Vilone and Luca Longo. "Explainable artificial intelligence: a systematic review". In: *arXiv preprint arXiv:2006.00093* (2020) (cit. on p. 17).

[319] Stefano Vincenzi, Angelo Porrello, Pietro Buzzega, et al. "The color out of space: learning self-supervised representations for earth observation imagery". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 3034–3041 (cit. on p. 15).

[320] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5797–5808 (cit. on p. 65).

[321] Johannes Wagner, Elisabeth Andre, Florian Lingenfelser, and Jonghwa Kim. "Exploring fusion methods for multimodal emotion recognition with missing data". In: *IEEE Transactions on Affective Computing* 2.4 (2011), pp. 206–218 (cit. on p. 16).

[322] Volker Walter. "Object-based classification of remote sensing data for change detection". In: *ISPRS Journal of photogrammetry and remote sensing* 58.3-4 (2004), pp. 225–238 (cit. on p. 13).

[323] Anna X Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. "Deep transfer learning for crop yield prediction with remote sensing data". In: *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 2018, pp. 1–5 (cit. on p. 15).

[324] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. "Extending global-local view alignment for self-supervised learning with remote sensing imagery". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 2443–2453 (cit. on p. 15).

[325] Björn Waske and Matthias Braun. "Classifier ensembles for land cover mapping using multitemporal SAR imagery". In: *ISPRS journal of photogrammetry and remote sensing* 64.5 (2009), pp. 450–457 (cit. on p. 13).

[326] Marie Weiss, Frédéric Jacob, and Grgory Duveiller. "Remote sensing for agricultural applications: A meta-review". In: *Remote sensing of environment* 236 (2020), p. 111402 (cit. on pp. 12, 29).

[327] RM Welch, SK Sengupta, AK Goroch, et al. "Polar cloud and surface classification using AVHRR imagery: An intercomparison of methods". In: *Journal of Applied Meteorology and Climatology* 31.5 (1992), pp. 405–420 (cit. on p. 13).

[328] Emily Hoffhine Wilson and Steven A Sader. "Detection of forest harvest type using multiple dates of Landsat TM imagery". In: *Remote Sensing of Environment* 80.3 (2002), pp. 385–396 (cit. on p. 124).

[329]Aleksandra Wolanin, Gonzalo Mateo-García, Gustau Camps-Valls, et al. "Estimating and Understanding Crop Yields with Explainable Deep Learning in the Indian Wheat Belt". In: *Environmental research letters* 15.2 (2020), p. 024019 (cit. on p. 45).

[330]Chaoyang Wu, Zheng Niu, Quan Tang, and Wenjiang Huang. "Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation". In: *Agricultural and forest meteorology* 148.8-9 (2008), pp. 1230–1241 (cit. on p. 142).

[331]Weiying Xie, Yuhang Cui, Yunsong Li, et al. "HPGAN: Hyperspectral pansharpening using 3-D generative adversarial networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.1 (2020), pp. 463–477 (cit. on p. 15).

[332]Cai Xu, Jiajun Si, Ziyu Guan, et al. "Reliable conflictive multi-view learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. 14. 2024, pp. 16129–16137 (cit. on p. 16).

[333]Jinfan Xu, Jie Yang, Xingguo Xiong, et al. "Towards interpreting multi-temporal deep learning models in crop mapping". In: *Remote Sensing of Environment* 264 (2021), p. 112599 (cit. on pp. 57, 68, 71).

[334]Wei Xu, Qili Wang, and Runyu Chen. "Spatio-temporal prediction of crop disease severity for agricultural emergency management based on recurrent neural networks". In: *GeoInformatica* 22 (2018), pp. 363–381 (cit. on p. 15).

[335]Jingyu Yang, Jianhua Guo, Huanjing Yue, et al. "CDnet: CNN-based cloud detection for remote sensing imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.8 (2019), pp. 6195–6211 (cit. on p. 15).

[336]Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. "On the (in) fidelity and sensitivity of explanations". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 25, 33, 71).

[337]Z. Yi, L. Jia, and Q. Chen. "Crop classification using multi-temporal Sentinel-2 data in the Shiyang River Basin of China". In: *Remote Sensing* 12.24 (2020), p. 4052 (cit. on p. 122).

[338]Yuan Yuan, Xiangtao Zheng, and Xiaoqiang Lu. "Hyperspectral image superresolution by transfer learning". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10.5 (2017), pp. 1963–1974 (cit. on p. 15).

[339]M Zehana, J Desachy, and EH Zahzah. "DEM features for remote sensing image analysis by an expert system: stream networks and real distances computation". In: [*Proceedings*] *IGARSS'91 Remote Sensing: Global Monitoring for Earth Management*. Vol. 3. IEEE. 1991, pp. 1861–1864 (cit. on p. 13).

[340]Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014*. Springer. 2014, pp. 818–833 (cit. on pp. 21, 31).

[341]Yelu Zeng, Dalei Hao, Alfredo Huete, et al. "Optical vegetation indices for monitoring terrestrial ecosystems globally". In: *Nature Reviews Earth & Environment* 3.7 (2022), pp. 477–493 (cit. on p. 119).

[342]Ce Zhang, Isabel Sargent, Xin Pan, et al. "An object-based convolutional neural network (OCNN) for urban land use classification". In: *Remote sensing of environment* 216 (2018), pp. 57–70 (cit. on p. 15).

[343] Kai Zhang, Xue Zhao, Feng Zhang, et al. "Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–15 (cit. on p. 15).

[344] Yu Zhang and Qiang Yang. "A survey on multi-task learning". In: *IEEE transactions on knowledge and data engineering* 34.12 (2021), pp. 5586–5609 (cit. on p. 83).

[345] Yuan Zhang, Fang Shen, Xuerong Sun, and Kun Tan. "Marine big data-driven ensemble learning for estimating global phytoplankton group composition over two decades (1997–2020)". In: *Remote Sensing of Environment* 294 (2023), p. 113596 (cit. on p. 106).

[346] Yuan Zhang, Bin Yang, Xiaohui Liu, and Cuizhen Wang. "Estimation of rice grain yield from dual-polarization Radarsat-2 SAR data by integrating a rice canopy scattering model and a genetic algorithm". In: *International journal of applied earth observation and geoinformation* 57 (2017), pp. 75–85 (cit. on p. 14).

[347] Bei Zhao, Bo Huang, and Yanfei Zhong. "Transfer learning with fully pretrained deep convolution networks for land-use classification". In: *IEEE Geoscience and Remote Sensing Letters* 14.9 (2017), pp. 1436–1440 (cit. on p. 15).

[348] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929 (cit. on pp. 22, 79).

[349] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Object detectors emerge in deep scene cnns". In: *International Conference on Learning Representations*. 2015 (cit. on p. 21).

[350] Min Zhou, Zhengxia Zou, Zhenwei Shi, Wen-Jun Zeng, and Jie Gui. "Local attention networks for occluded airplane detection in remote sensing images". In: *IEEE Geoscience and Remote Sensing Letters* 17.3 (2019), pp. 381–385 (cit. on p. 57).

[351] Yanan Zhou, Wei Wu, Huan Wang, et al. "Identification of soil texture classes under vegetation cover based on Sentinel-2 data with SVM and SHAP techniques". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 3758–3770 (cit. on p. 24).

[352] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. "Do feature attribution methods correctly attribute features?" In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9. 2022, pp. 9623–9633 (cit. on p. 19).

[353] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. "Visualizing deep neural network decisions: Prediction difference analysis". In: *arXiv preprint arXiv:1702.04595* (2017) (cit. on p. 21).

ACADEMIC CURRICULUM VITÆ: HIBA NAJJAR

EDUCATION

Ph.D. in Computer Science *2022-2025*
Title: "Enhancing Interpretable Machine Learning for Earth Observation", University of Kaiserslautern-Landau (RPTU), Germany.
Supervisor: Prof. Dr. Prof. h.c. Andreas Dengel.

M.Sc. in Applied Mathematics & Data Science *2018 - 2021*
Engineering School Mines Nancy, France.

EXPERIENCES

Research Assistant at DFKI - Germany *2022-2025*
Yield Consortium Project: agricultural yield prediction using temporal satellite data and advanced deep learning models.
Mission KI Project: development of uniform standards to ensure trustworthy deployment and usage of AI systems.

Data Science Intern at BASF Digital Farming - Germany *2021*
Development of semi-supervised learning networks for plant disease identification and image classification.

Junior Data Researcher at LORIA - France *2019-2020*
Revision of the generation process of the Protein-Protein Interactions Domain Miner (PPIDM) repository of inferred interactions between protein domains. (http://ppidm.loria.fr/)

CERTIFICATIONS

Machine Learning | Stanford University, by Andrew NG
Data Science Specialization | Johns Hopkins University
Deep Learning Specialization | deeplearning.ai
TensorFlow Developer Professional Certificate | deeplearning.ai
Natural Language Processing Specialization | deeplearning.ai
Generative Adversarial Networks (GANs) | deeplearning.ai