EXPLAIN IT TO ME - FACING REMOTE SENSING CHALLENGES IN THE BIO- AND GEOSCIENCES WITH EXPLAINABLE MACHINE LEARNING

R. Roscher[1,2,*], B.Bohn[3], M. F. Duarte[4], J. Garcke[3,5]

[1] Institute of Geodesy and Geoinformation, University of Bonn, Germany - ribana.roscher@uni-bonn.de
[2] Institute of Computer Science, University of Osnabrueck, Germany
[3] Institute for Numerical Simulation, University of Bonn, Germany - bohn@ins.uni-bonn.de
[4] Department of Electrical and Computer Engineering, University of Massachusetts Amherst, USA - mduarte@ecs.umass.edu
[5] Fraunhofer Center for Machine Learning and Fraunhofer SCAI, Sankt Augustin, Germany - jochen.garcke@scai.fraunhofer.de

KEY WORDS: Machine Learning, Explainability, Interpretability, Geosciences, Biosciences

ABSTRACT:

For some time now, machine learning methods have been indispensable in many application areas. Especially with the recent development of efficient neural networks, these methods are increasingly used in the sciences to obtain scientific outcomes from observational or simulated data. Besides a high accuracy, a desired goal is to learn explainable models. In order to reach this goal and obtain explanation, knowledge from the respective domain is necessary, which can be integrated into the model or applied post-hoc. We discuss explainable machine learning approaches which are used to tackle common challenges in the bio- and geosciences, such as limited amount of labeled data or the provision of reliable and scientific consistent results. We show that recent advances in machine learning to enhance transparency, interpretability, and explainability are helpful in overcoming these challenges.

1. INTRODUCTION

The usage of machine learning (ML) methods, especially neural networks (NN), for scientific applications has grown considerably in the last years. A major difference between commercial and scientific applications is that ML models are not only trained with regard to high accuracy, but there is also a high demand for understanding the way that a specific model operates and the underlying reasons for the produced decisions. This is contrary to the black box behaviour of ML methods, and implies that *explainable machine learning* approaches are required to solve application-specific tasks.

In the bio- and geosciences, where remote sensing in combination with ML has proven to be an indispensable means for an efficient provision of scientific results, the general goals for utilizing ML are scientific understanding, inferring causal relationships from observational data, or even achieving new scientific insights. The reasons for demanding a higher level of explanation for machine learning models are diverse and vary depending on the application and the user's intentions, purposes, goals or contextual accuracy standards. In the broader context, properties that can be relevant when considering explainability of ML algorithms are safety/trust, accountability, reproducibility, transferability, robustness and multi-objective trade-off or mismatched objectives, see e.g. (Doshi-Velez, Kim, 2017, Lipton, 2018). In the specific context of remote sensing applications, explainable machine learning can be used to tackle common challenges. We identified four of them, as illustrated in Figure 1, starting from the specific characteristics of remote sensing data and reaching to the cross-domain challenge of providing a transparent, interpretable, and explainable model, but also the challenge of deriving scientific outcomes from output results. Although many classical ML approaches are already employing basic interpretability and explanation concepts such as feature detection mechanisms or output visualization tools, the success of NN models leads to the necessity of adapting and extending these concepts to novel approaches and of formulating more sophisticated explanation schemes.

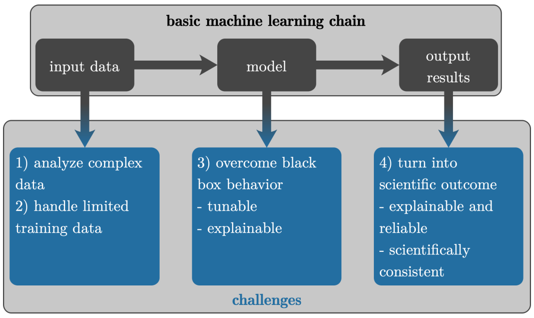
basic machine learning chain
input data
model
output results
1) analyze complex data
2) handle limited training data
3) overcome black box behavior
- tunable
- explainable
4) turn into scientific outcome
- explainable and reliable
- scientifically consistent
challenges

Figure 1. Basic machine learning chain to derive input-output relations. Various challenges arise when a machine learning method is applied in remote sensing.

In this paper, we focus on remote sensing applications in the area of the bio- and geosciences such as satellite-based analyses of processes and phenomena related to the Earth, but also analysis of close-range sensor data in the context of plant sciences. The main contribution of this paper is the discussion of explainable machine learning approaches which are used for these applications and especially those ones which showed to be beneficial to tackle common challenges from the remote sensing area. We continue our discussions from (Roscher et al., 2020) and consider the usefulness of explainable machine learning not only for deriving scientific output but also for tackling known remote sensing challenges.

2. FROM BLACK BOXES TO EXPLAINABLE MACHINE LEARNING MODELS

In the bio- and geosciences it is usually not enough to develop models that are optimized for accuracy. Other properties such

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-V-3-2020-817-2020

as transparency, interpretability and the possibility of integrating domain knowledge are also desirable. Especially deep neural networks, which provide high accuracy in many application areas, are said to lack these properties.

The literature provides several terms which are related to explainable machine learning with only partially consistent meanings, see e.g. (Doshi-Velez, Kim, 2017, Gilpin et al., 2018, Guidotti et al., 2018, Lipton, 2018, Montavon et al., 2018, Murdoch et al., 2019). For this work, we rely on the definitions given in (Roscher et al., 2020) and discuss them with regard to applications in the bio- and geosciences. In contrast to recent works which focus on the integration of domain knowledge to enhance the scientific consistency and plausibility (von Rueden et al., 2020, Karpatne et al., 2017), we focus on domain knowledge that is used for explainable machine learning.

2.1 Transparency

Transparency can be considered as the first step to open black box models. Generally, an ML approach is transparent if the training and testing process can be described and motivated by the approach designer. Here, transparency is connected to various aspects such as the overall model structure, model components or the learning algorithm. One can differentiate even more precisely between three kinds of transparency that involve different components: model transparency, design transparency, and algorithmic transparency. An often mentioned example of a non-transparent model, i.e. a black box model, is a neural network. However, the model can be considered transparent as soon as its input-output relation and structure can be written down in mathematical terms, which is generally the case if the model architecture and hyperparameters are given. Note that three outlined transparency aspects of the model also relate to the topic of reproducibility. Without a reproducible prediction modelling pipeline and a reproducible outcome from a machine learning approach, it is unlikely that we can obtain a useful explanation of the results.

Domain knowledge can often boost design transparency. For example, individual model components such as single modules can be chosen based on knowledge from the specific application domain. Hyperparameters, on the other side, are not linked to domain knowledge but rather chosen heuristically and non-transparent from an algorithmic viewpoint.

2.2 Interpretability

As stated by (Montavon et al., 2018), an interpretation can be seen as a mapping of an abstract concept, such as ML learning behavior, into a domain that the human can make sense of. Several interpretation tools that help to understand ML models and their decisions better have been proposed recently. Further details, types of interpretation, and specific realization can be found in recent surveys (Adadi, Berrada, 2018, Gilpin et al., 2018, Guidotti et al., 2018). Visualization techniques that produce heatmaps showing relevant patterns in the input given the learned model are commonly used. Among these approaches are saliency/sensitivity maps, attention maps, relevance maps, or feature importance maps (Hohman et al., 2018, Montavon et al., 2018, Olah et al., 2018, Fukui et al., 2019). Note that these approaches reveal different properties on how the inputs affect the ML model, e.g., how much does an attribute contribute to a prediction versus how strongly does the prediction change when an attribute is changed. Therefore, these maps lead to different interpretations and explanations.

Other approaches use proxies that approximate complex models, such as neural networks, with simpler and more interpretable designs such as decision trees or linear models (Gilpin et al., 2018, Guidotti et al., 2018). Prototype selection produces archetypal or representative samples that summarize the input data given the ML model and its decisions, usually based on similarity of the input for the same prediction. In the case of unlabelled data, linear or nonlinear dimensionality reduction can be used to analyze or visualize the data, and the obtained low-dimensional embedding or latent variable model can be interpreted (Lee, Verleysen, 2007, Cichocki et al., 2009), e.g., by comparing to (known) physical quantities.

2.3 Explainability

(Guidotti et al., 2018) write that the definitions of explanations in AI assume implicitly that the concepts expressed in the understandable terms composing an explanation are self-contained and do not need further explanations. Therefore, it is usually proposed to achieve explainability purely algorithmically. On the other hand, we expect that domain knowledge is needed to infer explanations suitable for the underlying scientific investigation. Furthermore, the needed explanations will depend on the type of scientific analysis. Getting a first insight into which attributes influence an ML prediction at all is a different goal than investigating why example A shows a different behavior than example B. Overall, the scientist is using the data, interpretations, and domain knowledge to explain the output results (or the data) in correspondence to the scientific goal.

Note that the interpretation of a model — in understandable terms to a human — for an individual datum, on its own, might not provide an explanation to understand the decision. For example, the most relevant variables might be the same for several data, but the important observation for an understanding of the overall predictive behavior could be that in a ranking with respect to the interpretation, different variable lists are determined for each datum as being of relevance.

2.4 The role of domain knowledge

In traditional machine learning, domain knowledge is used for feature engineering and data preprocessing. We consider two additional major reasons why the integration of domain knowledge is an indispensable part of a successful machine learning model. First, domain knowledge is an essential part of explainability. As pointed out in the previous section, we state that a machine learning model cannot be explained without domain knowledge. Domain knowledge can be integrated into the model in different ways and help to achieve a higher level of explainability. It can also be applied *post-hoc*, e.g., to the model components or the output in order to explain them.

As second reason, we consider domain knowledge to be beneficial for restricting the solution space to scientifically consistent and plausible solutions. Not only does this lead to more scientifically valuable and more trustworthy results, but it can also help with the treatment of small data scenarios, or be used for performance reasons. (Reichstein et al., 2019) identify scientific consistency besides interpretability as one of the five major challenges that must be tackled in order to successfully adopt deep learning approaches in the geosciences. Note however that the explicit restriction of the solution space to scientifically consistent and plausible solutions is not a requirement to achieve explainable models and results and valuable scientific outcomes. Nonetheless, neglecting this restriction means that a consistent and plausible solution cannot be guaranteed, even if an optimal result has been achieved from a mathematical point of view.

(Karpatne et al., 2017) term the integration of physical knowledge into machine learning methods *theory-guided data science* and discuss several applications of it. Additionally, (Raissi et al., 2017) consider this aspect with a special focus in their work on physics-informed learning. A general taxonomy for an explicit integration of scientific knowledge into the ML pipeline, called *informed ML*, is proposed by (von Rueden et al., 2020). Their work states that three aspects are involved in an informed machine learning process: the type of integrated knowledge, the representation and transformation of knowledge to be integrable, and the location where the knowledge is integrated into the ML approach. Moreover, they point out that knowledge can be integrated into the data, the model learning process, or in the patterns. Generally, they point out that scientific consistency at the design stage can be understood as the result of a regularization effect, where various ways exist to restrict the solution space to scientifically consistent solutions.

Different types of domain knowledge from the bio- and geosciences can be integrated, where scientific consistency can be considered *a priori* at the model design stage or *a posteriori* by analysing the output results. For example, in the environmental and Earth system sciences, knowledge is oftentimes governed by scientific laws, analytic expressions, or differential equations (Reichstein et al., 2019). Knowledge can also be exploited by numerical simulations (Moseley et al., 2019), or by considering invariance, e.g., to scale (Murray et al., 2019) or to rotations (Cheng et al., 2016), or by encoding equivariance within the structure of a model. A popular method, which is mainly used for neural networks, is pre-training with simulation data and fine-tuning with actual labeled data from the application, which is usually limited.

Remote sensing data in general is geolocated, meaning that geospatial information such as coordinates are given along with actual sensor data. Due to the known position of the acquired data and meta-information such as the characteristics of the area (e.g., agricultural area, urban area, etc.) and the application task (e.g., building detection, biomass estimation, etc.), a variety of prior information can be integrated. For example, tasks related to man-made objects can include many different geometric properties, the assumption about uniform object sizes in metric space, and relationships between instances and/or classes. These can be represented in form of rules or constraints, ontologies, symmetries, or similarity measures.

Another valuable source of information is world knowledge, i.e., knowledge which is generally known to humans. For example, knowledge about land use, land cover changes, and (especially) their transition probabilities can be utilized. Thus, a transition from forest to burnt area might have higher likelihood than in the opposite direction, and it is obvious that desert areas tend not to turn into water areas. On the other hand, with knowledge of the phenological stages of plants and prior knowledge of crop rotations, highly probable transitions can be defined in advance. Another kind of knowledge that is much harder to integrate because it is not formalized is experts' intuition and knowledge from experience, given that such knowledge is difficult to represent as something that the computer can interpret. Here, human interaction is necessary to transform the given information, whereby the quality of the chosen representation is difficult to evaluate.

3. EXPLAINABLE MACHINE LEARNING FOR REMOTE SENSING

In recent years, the possibilities for increasing the explanatory power of machine learning models have grown. Although the motivations are similar in different communities, they differ based on the application and the user. This section reviews several applications from the bio- and geosciences that are addressed by remote sensing techniques and explainable machine learning approaches. Note that this collection of research works is a non-exhaustive selection from the most recent years of the literature. For each application, we start with a characterization of the challenge in more detail, followed by relating this to long-established approaches. We also look at novel approaches that have emerged mainly due to the uptake of NNs in this area, as they are the dominating ML approach used in the recent literature.

3.1 Challenge: Complex Data

Remote sensing data is oftentimes characterized by multimodality, high dimensions (e.g., when using hyperspectral sensors), and various noise sources. As a result, patterns in the data and relations are difficult to discover, and preprocessing is necessary in the course of exploratory data analysis.

A long-standing and widely used method to transfer acquired data into a more interpretable data representation is feature extraction. For instance, (Laparra et al., 2015) use an iterative dimensionality reduction approach based on principal component analysis for hyperspectral image data in order to determine which components contribute the most to reconstructing the original data and manage to accurately retrieve temperature and water vapor parameters. Although approaches such as feature selection, visualization techniques, or extraction of prototypical examples are nothing new, they are still essential and valid research tools to better understand the behaviour of models and their decisions, and to explain them in the context of the specific application domain. However, if recent approaches such as NNs are used, classical interpretation tools may not be applicable or do not lead to the desired success, and therefore new interpretation tools need to be developed and existing ones need to be improved and adapted. It should also be noted that although the approaches mentioned above may provide better interpretability, there is no guarantee of an explanation. For example, if the underlying factors of variation are not captured, explanations may not be possible or may even be misleading.

As a modern approach used in neural networks, (Lorenzo et al., 2018) use attention modules in CNNs to select bands from hyperspectral images that contain the most important information for a given task. The attention modules are inserted in various depths in the network and output heatmaps, which identify the most informative regions in different scales. Confidence scores are computed from the heatmaps and the classifier's output. Combining these with an elliptic envelope algorithm, the most meaningful bands are selected which are responsible for distinguishing land cover classes. Moreover, they analyze the impact of using attention modules and can underline that they do not influence the classification accuracy. Therefore, attention modules can be used to enhance interpretability and explainability without decreasing the accuracy. With the same goal, (Nagasubramanian et al., 2019) use saliency maps obtained from 3D CNNs meant for hyperspectral plant disease classification to provide physiological insights into the network predictions. In detail, they use the maps to identify the hyperspectral wavelengths and pixel locations that contribute most to the classification result. Both saliency maps in the spatial and spectral domain are visualizations understandable by a domain expert and contribute significantly to the result's confidence.

(McGovern et al., 2019) introduce composite saliency maps for tornado prediction based on radar images. By generating a human-understandable composite of saliency maps for all input images, they are able to explain input-output relations such as the probability of tornado events given a specific meteorological scenario. Their work shows that saliency maps reflect meaningful relationships that are learned by the model and that are in compliance with concepts familiar to meteorologists.

With the aim of obtaining a better understanding of complex remote sensing observations, (Wolanin et al., 2020) perform wheat yield estimation for the Indian Wheat Belt based on MODIS products and meteorological data. They utilize regression activation mapping, which is especially suitable for interpreting time-series data, for deeper analyzing their CNN. These activation mappings are meant to localize important patterns in time-series inputs and support findings of the network predictions. (Wolanin et al., 2020) show that the found patterns can be explained with specific domain knowledge and that they can be related to crop growth and meteorological conditions. This underlines the assumption that such interpretation tools can lead to novel insights about influencing features and events.

Another interpretation tool that provides relevant and human understandable information about the input is the summarization of the input data to so-called prototypes, which represent the patterns the network considers as most associated with the output. (Toms et al., 2019) apply this technique to an NN model which was trained for the analysis of climate patterns, where their prototypes are named optimal inputs. In their considered scenario they assign sea surface temperature anomaly maps to *El Niño* and *La Niña* events, where protoypes for both outputs were produced. Based on the prototypes, relevant spatial structures that are typical for both effects can be observed. Although this states a simple application scenario, it shows that prototypes support the user to quantify patterns in the data that maximize the envisioned outcome of the network.

3.2 Challenge: Limited Amount of Data Labels

A general challenge in remote sensing is that labeled samples are scarce. The reason is that manually labeling data is expensive and time-consuming, and even sometimes unbearable or impossible, e.g., if the study area is not accessible. Moreover, the collection of labeled samples is subjective, which can result in a biased model. Explainable machine learning models, once trained and in combination with interpretation tools, can be used to identify new samples that are valuable to increase stability, robustness, and the generalization ability of the current model, or for model transfer to changing environments.

A long-standing approach to this challenge is self-training, where the training data set is extended by samples that are considered relevant after application of the current model, and active learning, where additional training samples are acquired through the guidance of a user (Tuia et al., 2009). Self-training is used to avoid the creation or usage of large sample sets of labeled data, i.e. to adaptively enhance the available data set by samples which contribute most to the ML model's error reduction for instance. For example, (Baetens et al., 2019) use active learning and random forests for cloud detection and the creation of visualization masks in multispectral observations. Their approach consists of determining pixels or regions of low classification confidence in each ML step and retraining only in the respective areas. In this way, a large computational overhead is avoided and the user has a means of directly forcing the ML model to reiterate on obvious misclassifications. Considering scale invariance or equivariance within a model has been shown to improve performance for tasks where training data is limited and the number of model parameters must be kept low (Murray et al., 2019). Such approaches have always been very common and are also used for neural networks.

(Bi et al., 2019) employ both data augmentation methods and active learning mechanisms to remedy the problem of scarce amounts of labeled input data for polarimetric synthetic aperture radar image classification. They train a multi-layer CNN by using only a small amount of annotated input pixels. Then, the active learning decision mechanism considers ambiguous samples (i.e., samples for which their probabilities of belonging to the two most probable classes are close to each other) as inputs for the next ML iteration. Furthermore, invariance inspired data augmentation is performed via image patch rotation and flipping. Afterwards, the model is retrained and the process is repeated until the prediction error is sufficiently small. Finally, energy minimization over a graph-based Markov random field model is used to achieve noise reduction and spatial consistency of the NN predictions.

With similar ideas, so-called adversarial examples are exploited. They are caused by artificial or unforeseen disturbances and are oftentimes imperceptible by humans. They are known to deteriorate the accuracy and stability of NNs (Chen et al., 2019). Adversarial attacks can be seen as an interpretation tool, given that they help to understand the weaknesses of the trained NN model and the shortcomings of the given training data set. As stated by (Goodfellow et al., 2014), training on adversarial examples can regularize the model. Therefore, the identification of such samples can be beneficial for model training. For example, (Yang et al., 2019) utilize CNNs and class activation maps (Grad-CAM++ , (Chattopadhay et al., 2018)) for land use classification and object detection to identify regions in aerial images which contribute most to the networks' decision. By training the network with images in which the most contributing regions are occluded on purpose (i.e., adversarial examples), the network additionally learns on samples which are difficult to predict. Their motivation is to enhance robustness and the generalization ability of the network by this manipulating technique, which is measured by the increase of the accuracy on the test set. Despite the success of active learning and adversarial attacks, they differ from explainable machine learning approaches where domain knowledge is exhaustively exploited to guide the learning process rather than automatically generated evaluation criteria such as classifier's confidence.

(Janik et al., 2019) point out that explainable models help to identify properties of the data which result in good model performance. This information helps to improve the model itself and guides a better selection of the input data. In many remote sensing applications where in-situ data is needed as training data, this time-consuming collection process can be optimized with the knowledge about the value of specific data samples. In their work, they introduce an interactive visualization tool that uses learned latent representations and an intersection over union values. The latent representation is not inherently understandable by a human, but can be visualized by PCA. Given the low-dimensional representation and the intersection over union values, image patches with similar values can be grouped together regarding important features for the network. One motivation behind this tool is a pre-screening of potential image patches that are valuable for labeling.

3.3 Challenge: Overcome Model's Black-Box Behavior

Remote sensing applications usually involve very complex processes and phenomena. Therefore, approaches such as neural networks are promising because of their ability to characterize complex relationships. However, the supposed black-box behavior of these methods leads various communities to resist their use. One challenge is the model tuning, e.g., the choice of hyperparameters or the adaptation of the design regarding a specific application. Due to the complexity of the model and a general lack of interpretability, the model's improvement is not trivial and involves time-consuming search processes. So far this challenge has received little attention, as models have been used that are easier to understand, such as decision trees. However, with the use of complex neural networks this point has become one of the main challenges in remote sensing.

Besides the visualization of single model components (e.g., single filters in neural networks to evaluate the learned structures (Marcos et al., 2018, Gagne II et al., 2019)), additional interpretations tools are used to understand the model behavior. Class activation maps (CAMs) are used by (Fu et al., 2019) in their MultiCAM approach. An NN consisting of two subnetworks is used for aircraft detection in aerial images. The first sub-network extracts features from the whole image and delivers object saliency maps for each class, whereas the second subnetwork learns features from a focused region that is derived from a combination of all class-wise saliency maps. Classification is performed on fused features from both sub-networks. In contrast to other works using CAM (e.g., (Vasu et al., 2018)), this approach uses the saliency maps of all classes and combines them into one. In this way, the object will be localized even if classified incorrectly. Overall, their procedure can thus not only help to identify the part of the image that is the focus of the network, but also improves model performance due to a more focused learning procedure.

The interpretable and explainable LSTM network presented by (Hochreiter, Klambauer, 2019) for rainfall-runoff forecasting is one of the first works in the geosciences where single model components are interpreted and explained. The general idea is to relate the model components to specific application scenarios, system parameters, or states. They succesfully analyze whether single components of the trained model are in compliance with hydrological processes. In this case, they focus on memory cells, which are a crucial element of LSTMs, and correlate the cell state with hydrological storage processes. In this way, they relate single memory cells of the LSTM to hydrological system states such as snow. Moreover, interactions between neurons turn out to be useful indicators for influences from meteorological input variables such as temperature. Although the results are preliminary, this can be seen as one of the first promising attempts to inform domain experts, e.g., hydrologists, about patterns and relations the network found. In this way, the network can be optimized in a significantly more intuitive way than by using common hyperparameter and architecture searches. With a similar idea, (Marcos et al., 2019) use semantically interpretable activation maps, an interpretable intermediate representation in a neural network to better understand which features distinguish scenic and non-scenic images. (Camps-Valls et al., 2018) use physics-aware Gaussian processes with the goal of estimating bio-physical parameters, e.g., leaf area index from remote sensing data. For this, they learn a latent force model with incorporated ordinary differential equations and interpret it in view of the physical mechanism that generated the input-output relations. In their work, they show that one latent function captured the smooth and periodic component of the output, while two others focus on the noisier part with an important residual periodical component.

An interactive concept called explanatory interactive learning is introduced by (Schramowski et al., 2020) with the goal to improve the model using corrected explanations. In their work, sugar beet plants are classified as diseased or healthy and the classified result is provided to an expert along with class activation heatmaps. Interpretations that do not fit to the human explanation, but were classified correctly, are identified and formulated as penalty terms. In this way, the human expert in the loop constrains the interpretations provided by the learned model to match domain knowledge. An improved model can be derived yielding more reasonable interpretations and thus help form better explanations and more reliable results. Similar ideas have been proposed, e.g., by (Ghosal et al., 2018), where they use interpretation tools for image-based plant stress phenotyping. They produce so-called explanation maps as sums of the most important features maps indicated by their activation level to isolate the visual cues for stress and disease symptoms. However, their work focuses on the comparison of manually marked visual cues by an expert and the automatically derived explanation maps, rather then the improvement of the model and its interpretations.

One of the landmark problems in hyperspectral imaging and remote sensing is spectral unmixing: given that a single pixel in a remote sensing hyperspectral image will correspond to multiple elements or types of surface, it is desirable to identify the individual types of surface or elements present in the pixel. Several methods from ML have been leveraged to solve the unmixing problem (Heylen et al., 2014), including supervised learning methods (e.g., SVM, NN) and unsupervised learning methods (e.g., linear regression). Among these, linear regression arguably provides the most transparent approach to the problem, since it seeks the simplest linear combination of base element spectra that matches the observation. However, the physical process involving the spectrum acquisition from the mixture does not follow the linear model assumed by the regression methods, and nonlinear unmixing approaches have been proposed to remediate this discrepancy (Dobigeon et al., 2014).

Methods that move further into the direction of explainable ML for spectral unmixing increase the accuracy of the mixture model by enforcing the physical behaviors underlying the acquisition (Close et al., 2012). A simple approach learns a map from the true nonlinear mixture to the idealized linear mixture (e.g., learned by a NN) to which regression is applied (Koirala et al., 2018). However, approaches like this leverage the power of black-box models without any underlying intuition as to the accuracy of the model; e.g., there is no explainable characterization of the nonlinear mixture-to-linear mixture mapping involved. A variety of approaches to unmixing that address the nonlinear nature of the mixture leverage the Hapke mixture model, which provides an approximation of the geometrical optics underlying light scattering (Hapke, 1963, Shkuratov et al., 2012). In particular, several efforts proposed methods based on NNs that learn the Hapke mixing equation as a nonlinear mapping from spectra to mixture element concentrations (Guilfoyle et al., 2001, Altmann et al., 2011). Other methods use kernel-based factorizations that integrate the nonlinear model in the chosen kernel (Wang, Qian, 2016).

3.4 Challenge: Turn Results into Scientific Valuable Outcomes

Related to the aforementioned challenges, another reason to strive for explainable ML models is to obtain scientific insights and novel discoveries (Roscher et al., 2020). Many works address the task to learn input-output relations from observations and transfer them to new input-output relations, given the learned model. However, this does not imply that the outcome is valuable from a scientific point of view. The explainability of the results is higher if, for example, derived parameters are known and usable in the application domain (e.g., as input for further simulations) rather than naive outcomes from which it is possible to deduce what happens, but not why. Thus, learned models and tools which help to derive the actual reasons of the occurrence of the outcome foster the scientific value of it.

Scientific insights and novel discoveries can only be obtained if the outcome is reliable, scientifically consistent and plausible, and can be exploited by the user. A major point is that the reliability of the result can be questioned, even with given confidence scores; therefore, trust in the result is not given. One possible reason for such questioning is that an overly complex model is not comprehensible to a human. However, interpretation tools such as those presented in (McGovern et al., 2019, Hochreiter, Klambauer, 2019, Wolanin et al., 2020, Toms et al., 2019, Nagasubramanian et al., 2019) can be used to verify the outcome using domain knowledge, and if necessary the model can be improved as presented in (Schramowski et al., 2020) and (Kierdorf et al., 2020). Methods that enhance or enforce scientific consistency and plausibility also play an important role, as presented in (von Rueden et al., 2020).

4. CONCLUSION AND FUTURE DIRECTIONS

In this work we discussed how explainable machine learning can be beneficial when tackling commonly known challenges in remote sensing for the bio- and geosciences. We focus on identifying challenges where explainable machine learning can contribute; nonetheless, we also see other promising directions. This survey intended to review current advances towards explainable machine learning in the bio- and geosciences, and to provide new ideas and research directions, but also to increase interest in novel machine learning algorithms.

Although related research directions such as hybrid learning (Reichstein et al., 2019) and informed machine learning (von Rueden et al., 2020) are not covered in detail in this paper, we want to point these out as important future directions. In the broadest sense, various post-processing steps that are already widely used can be seen as informed machine learning approaches. However, these approaches are generally performed in two steps, and therefore finding an optimal solution with respect to given domain knowledge is not guaranteed. With the development of efficient optimization strategies that go hand in hand with the rapid development of neural networks, we face new possibilities to integrate domain knowledge early during the learning process and to optimize the model accordingly.

We are convinced that incorporating domain knowledge to gain explainability is a crucial next step in enabling explainability for ML in remote sensing applications. To this end, the error measure according to which the NN model is optimized should take explanation errors into account, see (Rieger et al., 2019, Schramowski et al., 2020). For instance, if certain parts of a hyperspectral image do not contribute much to the quantity of interest, they should be weighted less by the NN when making its decision. Furthermore, when it comes to analyzing complex data and gaining transparency for NN approaches a statistical interaction analysis of the network's degrees of freedom can help to identify the most important contributions of both the input features and the network weights, see (Tsang et al., 2018).

Another promising direction are domain-specific design choices of the machine learning model (Beucler et al., 2019, Camps-Valls et al., 2018). For example, neural networks offer a variety of possibilities ranging from a specific choice of the architecture to the integration of modules. As presented by (Iten et al., 2020), single model components such as neurons can be used to capture disentangled factors of variations. Additionally, interactions between features can be represented by specific neural network architectures (Tsang et al., 2018). Furthermore, specific modules can be incorporated such as grouping layers (Yan et al., 2019) or attention modules (Lorenzo et al., 2018), which provide added value in terms of interpretability and thus the possibility of explanation, but do not necessarily have to significantly influence the task-solving process.

A future direction which has barely been considered so far in this context are evaluation metrics related to explainability. For scenarios with small sample size or in view of adversarial attacks, high accuracies are not inherently an indicator for confidence and reliability. Moreover, accuracy does not specify the quality of the outcome from a scientific viewpoint. Therefore, additional evaluation criteria need to be developed. For example, (Karpatne et al., 2017) introduce a scientific inconsistency value for an informed machine learning method defined as a comparison between estimated test values and values calculated from known physical laws. (Kailkhura et al., 2019) propose a measure of trustworthiness of classification predictions taking into account the average Gower distance from a test sample to labeled samples in the same class and the average Gower distance to labeled samples of other classes. (Lapuschkin et al., 2019) analyze the predictions of a model semi-automatically in order to check the reliability of the obtained results. They cluster heatmaps obtained from spectral relevance analysis and visually inspect the content of the cluster. By that they uncover spurious behavior of a model such as the clever-Hans-effect, where right decisions are taken for wrong reasons. This has also been successfully applied to hyperspectral plant disease detection (Schramowski et al., 2020).

Finally, we want to emphasize that *causality* will probably be one of the most important steps in future MLf research in relation to explainable AI. See (Runge et al., 2019) for an overview of causal inference approaches in Earth system sciences and (Perez-Suay, Camps-Valls, 2019) for a study on observation-based causal inference in remote sensing and geosciences.

REFERENCES

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.

Altmann, Y., Dobigeon, N., Tourneret, J., McLaughlin, S., 2011. Nonlinear unmixing of hyperspectral images using radial basis functions and orthogonal least squares. *IEEE IGARSS*, 1151–1154.

Baetens, L., Desjardins, C., Hagolle, O., 2019. Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure. *Remote Sensing*, 11(4), 433.

Beucler, T., Pritchard, M., Rasp, S., Gentine, P., Ott, J., Baldi, P., 2019. Enforcing Analytic Constraints in Neural-Networks Emulating Physical Systems. *arXiv:1909.00912*.

Bi, H., Xu, F., Wei, Z., Xue, Y., Xu, Z., 2019. An Active Deep Learning Approach for Minimally Supervised PolSAR Image Classification. *IEEE Trans Geosci Remote Sens*, 57(11), 9378–9395.

Camps-Valls, G., Martino, L., Svendsen, D. H., Campos-Taberner, M., Muñoz-Marí, J., Laparra, V., Luengo, D., García-Haro, F. J., 2018. Physics-aware Gaussian processes in remote sensing. *Applied Soft Computing*, 68, 69–82.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V. N., 2018. Grad-Cam++: Generalized gradient-based visual explanations for deep convolutional networks. *IEEE Winter Conference on Applications of Computer Vision*, 839–847.

Chen, L., Zhu, G., Li, Q., Li, H., 2019. Adversarial Example in Remote Sensing Image Recognition. *arXiv:1910.13222*.

Cheng, G., Zhou, P., Han, J., 2016. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), 7405-7415.

Cichocki, A., Zdunek, R., Phan, A. H., Amari, S. I., 2009. *Nonnegative Matrix and Tensor Factorization*. John Wiley & Sons.

Close, R., Gader, P., Wilson, J., Zare, A., 2012. Using physics-based macroscopic and microscopic mixture models for hyperspectral pixel unmixing. *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII*, 8390, SPIE, 469 – 481.

Dobigeon, N., Tourneret, J., Richard, C., Bermudez, J. C. M., McLaughlin, S., Hero, A. O., 2014. Nonlinear Unmixing of Hyperspectral Images: Models and Algorithms. *IEEE Signal Processing Magazine*, 31(1), 82-94.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*.

Fu, K., Dai, W., Zhang, Y., Wang, Z., Yan, M., Sun, X., 2019. Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images. *Remote Sensing*, 11(5), 544.

Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H., 2019. Attention branch network: Learning of attention mechanism for visual explanation. *IEEE CVPR*, 10705–10714.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., Thompson, G., 2019. Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, 147(8), 2827–2845.

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., Sarkar, S., 2018. An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences*, 115(18), 4613–4618.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An overview of interpretability of machine learning. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89.

Goodfellow, I. J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. *arXiv:1412.6572*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 1–42.

Guilfoyle, K. J., Althouse, M. L., Chang, C.-I., 2001. A quantitative and comparative analysis of linear and nonlinear spectral mixture models using radial basis function neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 39(10), 2314-2318.

Hapke, B. W., 1963. A theoretical photometric function for the lunar surface. *Journal of Geophysical Research (1896-1977)*, 68(15), 4571-4586.

Heylen, R., Parente, M., Gader, P., 2014. A Review of Nonlinear Hyperspectral Unmixing Methods. *IEEE J Sel Top Appl Earth Obs Remote Sens*, 7(6), 1844-1868.

Hochreiter, S., Klambauer, G., 2019. NeuralHydrology–Interpreting LSTMs in Hydrology. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 11700, 347.

Hohman, F. M., Kahng, M., Pienta, R., Chau, D. H., 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 1–20.

Iten, R., Metger, T., Wilming, H., del Rio, L., Renner, R., 2020. Discovering Physical Concepts with Neural Networks. *Phys. Rev. Lett.*, 124, 010508.

Janik, A., Sankaran, K., Ortiz, A., 2019. Interpreting Black-Box Semantic Segmentation Models in Remote Sensing Applications. D. Archambault, I. Nabney, J. Peltonen (eds), *Machine Learning Methods in Visualisation for Big Data*.

Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A., Han, T., 2019. Reliable and Explainable Machine Learning Methods for Accelerated Material Discovery. *npj Comput. Mater.*, 5(108).

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2318–2331.

Kierdorf, J., Garcke, J., Behley, J., Cheeseman, T., Roscher, R., 2020. What identifies a whale by its fluke? On the benefit of interpretable machine learning for whale identification. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Science*. accepted.

Koirala, B., Heylen, R., Scheunders, P., 2018. A neural network method for nonlinear hyperspectral unmixing. *IEEE IGARSS*, 4233–4236.

Laparra, V., Malo, J., Camps-Valls, G., 2015. Dimensionality Reduction via Regression in Hyperspectral Imagery. *IEEE Journal of Selected Topics in Signal Processing*, 9(6), 1026–1036.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), 1096.

Lee, J. A., Verleysen, M., 2007. *Nonlinear Dimensionality Reduction*. Information Science and Statistics, Springer.

Lipton, Z., 2018. The mythos of model interpretability. *Queue*, 16(3), 1-28.

Lorenzo, P. R., Tulczyjew, L., Marcinkiewicz, M., Nalepa, J., 2018. Band selection from hyperspectral images using attention-based convolutional neural networks. *arXiv:1811.02667*.

Marcos, D., Lobry, S., Tuia, D., 2019. Semantically interpretable activation maps: What-where-how explanations within CNNs. *Int. Conf. Computer Vision Workshop*, 4207–4215.

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, 96–107.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., Smith, T., 2019. Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11), 2175–2199.

Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing*, 73, 1-15.

Moseley, B., Nissen-Meyer, T., Markham, A., 2019. Deep learning for fast simulation of seismic waves in complex media. *Solid Earth Discussions*, 2019, 1–23.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.

Murray, J., Marcos, D., Tuia, D., 2019. Zoom in, zoom out: Injecting scale invariance into landuse classification CNNs. *IEEE IGARSS*, 5240–5243.

Nagasubramanian, K., Jones, S., Singh, A. K., Sarkar, S., Singh, A., Ganapathysubramanian, B., 2019. Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods*, 15(1), 98.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A., 2018. The Building Blocks of Interpretability. *Distill*. https://distill.pub/2018/building-blocks.

Perez-Suay, A., Camps-Valls, G., 2019. Causal Inference in Geoscience and Remote Sensing from Observational Data. *IEEE Trans Geosci Remote Sens*, 57(3), 1502–1513.

Raissi, M., Perdikaris, P., Karniadakis, G. E., 2017. Physics informed deep learning (Part II): data-driven discovery of nonlinear partial differential equations. *arXiv:1711.10566*.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. et al., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195.

Rieger, L., Singh, C., Murdoch, W. J., Yu, B., 2019. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv:1909.13584*.

Roscher, R., Bohn, B., Duarte, M. F., Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8, 42200–42216.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M., Muñoz-Marí, J., van Nes, E., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., Zscheischler, J., 2019. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10, 2553.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Luigs, H.-G., Mahlein, A.-K., Kersting, K., 2020. Right for the Wrong Scientific Reasons: Revising Deep Networks by Interacting with their Explanations. *arXiv:2001.05371*.

Shkuratov, Y., Kaydash, V., Korokhin, V., Velikodsky, Y., Petrov, D., Zubko, E., Stankevich, D., Videen, G., 2012. A critical assessment of the Hapke photometric model. *J. Quantitative Spectroscopy and Radiative Transfer*, 113(18), 2431 - 2456.

Toms, B. A., Barnes, E. A., Ebert-Uphoff, I., 2019. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *arXiv:1912.01752*.

Tsang, M., Cheng, D., Liu, Y., 2018. Detecting statistical interactions from neural network weights. *ICLR*.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., Emery, W. J., 2009. Active learning methods for remote sensing image classification. *IEEE Trans Geosci Remote Sens*, 47(7), 2218–2232.

Vasu, B., Rahman, F. U., Savakis, A., 2018. Aerial-Cam: Salient structures and textures in network class activation maps of aerial imagery. *2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, IEEE, 1–5.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., Schuecker, J., 2020. Informed machine learning - A taxonomy and survey of integrating knowledge into learning systems. *arXiv:1903.12394*.

Wang, W., Qian, Y., 2016. Kernel based sparse NMF algorithm for hyperspectral unmixing. *IEEE IGARSS*, 6970–6973.

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., Guanter, L., 2020. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters*, 15(2), 024019.

Yan, Y., Zhu, J., Duda, M., Solarz, E., Sripada, C., Koutra, D., 2019. GroupINN: Grouping-based interpretable neural network for classification of limited, noisy brain data. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 772–782.

Yang, R., Xu, X., Xu, Z., Ding, C., Pu, F., 2019. A class activation mapping guided adversarial training method for land-use classification and object detection. *IEEE IGARSS*, 9474–9477.