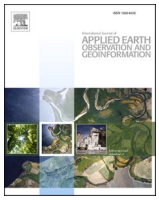
International Journal of Applied Earth Observations and Geoinformation 112 (2022) 102869

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Explainable AI for earth observation: A review including societal and regulatory perspectives

Caroline M. Gevaert

Dept. of Earth Observation Science, ITC, University of Twente, Enschede, the Netherlands

ARTICLE INFO

Keywords:
Earth observation
Remote sensing
Machine learning
Explainable artificial intelligence
Ethics
Regulations

ABSTRACT

Artificial intelligence and machine learning are ubiquitous in the domain of Earth Observation (EO) and Remote Sensing. Congruent to their success in the domain of computer vision, they have proven to obtain high accuracies for EO applications. Yet experts of EO should also consider the weaknesses of complex, machine-learning models before adopting them for specific applications. One such weakness is the lack of explainability of complex deep learning models. This paper reviews published examples of explainable ML or explainable AI in the field of Earth Observation. Explainability methods are classified as: intrinsic versus post-hoc, model-specific versus model-agnostic, and global versus local explanations and examples of each type are provided. This paper also identifies key explainability requirements identified the social sciences and upcoming regulatory recommendations from UNESCO Ethics of Artificial Intelligence and requirements from the EU draft Artificial Intelligence Act and analyzes whether these limitations are sufficiently addressed in the field of EO.

The findings indicate that there is a lack of clarity regarding which models can be considered interpretable or not. EO applications often utilize Random Forests as an "interpretable" benchmark algorithm to compare to complex deep-learning models even though social sciences clearly argue that large Random Forests cannot be considered as such. Secondly, most explanations target domain experts and not possible users of the algorithm, regulatory bodies, or those who might be affected by an algorithm's decisions. Finally, publications tend to simply provide explanations without testing the usefulness of the explanation by the intended audience. In light of these societal and regulatory considerations, a framework is provided to guide the selection of an appropriate machine learning algorithm based on the availability of simpler algorithms with a high predictive accuracy as well as the purpose and intended audience of the explanation.

1. Introduction

The past ten years have seen an incredible rise in the usage of Machine Learning (ML) and Artificial Intelligence (AI) in the domain of Earth Observation (EO) and Remote Sensing (RS). Indeed, there are now more than 1000 publications in the field (Camps-Valls et al., 2021; Zhu et al., 2017). Given the influence of these algorithms from the field of Computer Science on the domain of Earth Observation, it is perhaps wise to also consider their limitations. Indeed, increasing awareness of the fallacies of data-driven ML methods is calling for ethical guidelines for AI so society can responsibly utilize the great potential of these technologies. By 2020, the concepts of explainability and transparency were included in most guidelines on Responsible AI (Fjeld et al., 2020) as well as legislation. For example, the General Data Protection Regulation (GDPR) in Europe already demands a "right to explanation" (GDPR, 2016, Recital 71; Goodman and Flaxman, 2016) and the European Commission's draft Artificial Intelligence Act will implement transparency and explainability requirements for high-risk AI applications on the European Market (European Commission, 2021). Also in the geosciences, there is a clear call for evaluation frameworks to move away from assessing merely the performance of algorithms, but to also consider on the quality of the AI algorithms (Craglia, 2018).

The concept of explainability of machine learning (ML) models is not new. "Interpretable AI/ML" seems to be more prominent in the scientific community whereas "explainable AI/ML" is used more in a public setting (Adadi and Berrada, 2018). Research on the explainability of machine learning systems started as far back as the 1990s (Freitas, 2014; Holte, 1993). "Parsimony" in statistics and "simplification" in philosophy also describe the innate tension between increased accuracy of complex, data-driven models versus the interpretability and sometimes generalization capacity of simpler models (Herman, 2017). Still, since 2017 the terms of explainable and interpretable AI have gained

E-mail address: c.m.gevaert@utwente.nl.

https://doi.org/10.1016/j.jag.2022.102869
Received 12 April 2022; Received in revised form 30 May 2022; Accepted 12 June 2022
Available online 20 June 2022

momentum in published research (Arrieta et al., 2019).

If explainable ML is so important, then why is it not being implemented already? Burrell (2016) describes three main barriers for explainability: (1) intentional concealment of algorithms by institution (2) gaps in technical literacy, e.g. simply sharing the code isn't a sufficient explanation for most users, and (3) the mismatch between high-dimensional mathematical operations of state-of-the-art algorithms vs human-scale reasoning and styles of interpretation. I.e., it is extremely difficult for humans to understand the inner workings of very complex models, such as the popular deep learning models. Assuming they go coupled with adequate enforcement mechanisms, emerging regulatory frameworks may help overcome the first challenge. The field of explainable ML aims to overcome the second and third challenges. For example, by not only assessing the predictive accuracy of a ML model, but also the descriptive accuracy of the explainability method and the relevance of the explanation for the user Murdoch et al. (2019). Explanations of complex models don't replicate the exact reasoning of the model (Rudin, 2018), but rather attempt to distill the most influential factors behind the reasoning. Persuasive explanations can be distilled to be more convincing to the user (Herman, 2017) though increased abstraction comes at the cost of a loss of fidelity to the original model. Finding the balance between the complexity of deep learning models and how to simplify the inner workings to explanations understandable by humans is a tricky task which touches the fields of computer sciences and social sciences. The domain of Earth Observation is not escaping this drive for explainability. Developments in the field of Computer Science been successfully applied to the domain of Earth Observation, and it is to be expected that the drive for explainable algorithms will follow the same path. Indeed, explainability is highlighted as one of the six main research directions in the field (Tuia et al., 2021).

The objective of this paper is to consider the progress of explainable ML in the field of Earth Observation and highlight avenues for further research given the context of emerging regulatory requirements and known limitations of explainable ML methods from a social science perspective. It is not simply a review of explainable ML applications in Earth Observation, like Roscher et al., (2020a), but rather considers these developments in the broader societal context presented by social sciences and regulatory frameworks such as the UNESCO recommendations and draft European Artificial Intelligence Act.

The scope of this paper focusses on ML applications in Earth Observation and Remote Sensing. Typical tasks in this domain include the identification of objects in remotely sensed imagery or the classification of each pixel in the image (Ma et al., 2019; Zhu et al., 2017). As such, many of the ML tasks in this paper will be inspired by classification and semantic segmentation tasks in Computer Vision. This manuscript doesn't specifically consider other geospatial tasks which utilize other data types and may have specific data peculiarities, although there are likely to be many overlaps.

More specifically, this paper considers:

- Which explainable ML methods are being applied in the field of Earth Observation?
- Which key limitations of explainable ML are raised by the (upcoming) regulatory environment and social sciences?
- How can Earth Observation experts take these societal and regulatory concerns regarding explainable ML into account, to make sure they utilize and develop algorithms that align with these needs?

The paper is organized as follows. Section 2 addresses the ambiguity of the terms transparency, explainability, and interpretability and provides definitions and Section 3 provides an overview of the types of explainable ML methods. Section 4 describes the regulatory context regarding explainable ML and Section 5 describes concerns from the social sciences. This typology, regulatory context, and societal concerns are used as a frame to review the published research on explainable ML in the field of Earth Observation in Section 6. Section 7 discusses the findings and suggests a framework for Earth Observation scientists to use and Section 8 presents the conclusions.

2. Definitions

Interpretability doesn't have a formal technical meaning (Lipton, 2016), and there is much discrepancy in the use of terms such as transparency, interpretability and explainability in literature (Arrieta et al., 2019; Doshi-Velez et al., 2017; Hamon et al., 2020; Herman, 2017; Roscher et al., 2020b; Rudin, 2018; Sovrano et al., 2022). Transparency, usually refers to the model itself and how straightforward it is to access model parameters and motivate model decisions (Lipton, 2016; Roscher et al., 2020b; Tuia et al., 2021). As we will discuss below, regulatory frameworks often consider transparency to include the broader context of the ML system including the context, the availability of the code or datasets for auditing, and whether efforts were made to mitigate biases and ensure human rights (European Commission, 2021; UNESCO, 2021).

Many researchers opt to utilize the terms *explainability* and *interpretability* interchangeably (Du et al., 2018; Miller, 2019; Molnar, 2022) to signify "the degree to which an observer can understand the cause of a decision" (Biran and Cotton, 2017; Miller, 2019) or "how models are able to present reasonings in a way understandable to humans" (Du et al., 2018). However, others present a distinction where *interpretability* refers to the ability to understand how models reach certain predictions while *explainability* links such interpretations to domain knowledge (Roscher et al., 2020b; Tuia et al., 2021; Zhang et al., 2022). The main arguments for this distinction, which is prominent in the literature in the field of Earth Observation, is that this domain-specific contextual knowledge is required in order to interpret a model and that the goal of the user should be considered (Roscher et al., 2020b). For example, interpretability may help identify which features are more influential on a model's prediction, but explainability may incorporate domain knowledge to reason why these features are influential.

Finally, an *explanation* refers to how a model obtained a prediction for a single input sample (Miller, 2019; Molnar, 2022). The importance of domain knowledge on top of interpretability in order to achieve explainability is underscored by (Hamon et al., 2020), who state "just because the output of a model is interpretable doesn't mean that this interpretation is sufficient as an explanation, either considering the domain of application of the systems or from a legal point of view". Note that in an effort to reach human interpretability, *explanations, explainability* and *interpretability* often add components or simplifications to the system and it is therefore important to distinguish the results of these models from the original ML system (Doshi-Velez et al., 2017). The degree to which an explanation represents the full complexity of the predictive model is referred to as *model fidelity*.

3. Types of Explainable ML methods

Model transparency can be described at different levels. Roscher et al. (2020b) identify *model transparency* (similar to *simulatability* as described by (Lipton, 2016)), *design transparency,* and *algorithmic transparency*. Lipton, (2016) does not recognize design transparency but rather recognizes *decomposability* which refers to transparency at a parameter level. Interestingly, Roscher et al., (2020b) argue that transparency doesn't depend on the specific data, yet when considering the algorithm as a system and deployability to a new study area, the training data is very influential and indeed many regulatory frameworks consider a broader definition of transparency. We will come back to the consequences of neglecting this component later.

Interpretability and explainability methods can be grouped along different axes: intrinsic vs post-hoc, global vs local, and model-specific vs model-agnostic (Adadi and Berrada, 2018; Hamon et al., 2020; Lipton, 2016; Molnar, 2022; Murdoch et al., 2019; Schorr et al., 2021). Table 1 provides an overview of these categories. Note that intrinsic

Table 1
Overview of categories of explainability methods.

Design	**Intrinsic methods (I)** (a.k.a. model-based explanations) Explainability is integrated into the design of the algorithm and influence model predictions. 'Interpretable' or 'transparent' models fall into this category.	**Post-hoc methods (H)** Explanations are sought after the model has been trained and don't influence the model prediction.
Application	**Model-specific methods (MS)** The explainability method is tied to a specific type of ML model (e.g. neural networks). Intrinsic explanation methods are always model-specific.	**Model-agnostic methods (MA)** The explainability method can be relevant for many ML models.
Scope	**Global methods (G)** Describes the logic of the entire model.	**Local methods (L)** Provides an explanation for a single prediction.

explanation models are model-specific by nature (Adadi and Berrada, 2018). The boundaries between categories can be vague and also depend on how a model is applied. For example, ranking the most important features contributing the prediction of a model is generally a post-hoc explanation, but if the most important features from the ranking are utilized to retrain the model than this manuscript considers that explainability is intrinsically integrated into the ML workflow.

4. Regulatory context

The UNESCO Recommendation on the Ethics of Artificial Intelligence and the European Commission Draft Artificial Intelligence Act (AIA) are used to represent the regulatory context of explainable ML. The UNESCO Recommendations are selected for its global representativeness and emphasis on the context of Low- to Middle Income Countries (LMICs). The AIA is selected because the framework goes beyond recommendations and provides a legislative framework which enforces adherence. Moreover, if it follows the footsteps of the GDPR, it has the potential to form a global inspiration for other ML frameworks.

4.1. UNESCO

UNESCO adopted its "Recommendation on The Ethics of Artificial Intelligence" in November 2021 (UNESCO, 2021). It included an extensive peer-review process, with more than 800 responses for the online consultation and more than 500 participants at organized workshops. These Recommendations emphasize the importance of inclusion of LMICs and addressing digital and knowledge divides throughout the ML lifecycle. The Recommendation is non-binding, though UNESCO recommends Member States enforce these Recommendations and ensure the relevant parties, including the private sector, assume their responsibilities.

The Recommendations define values and principles that should guide the development and usage of ML systems. "Transparency and Explainability" is one of these principles. Transparency should promote the understanding of the ML system including the context, sensitivity, and any assurances to ensure safety or fairness. Explainability refers to "the understandability of the input, output and behavior of each algorithmic building block and how it contributes to the outcome of the systems". To ensure adherence to this principle, Member States are recommended to set requirements involving "the design and implementation of impact mechanisms" that consider the application, intended use, target audience and feasibility.

4.2. European Commission Artificial Intelligence Act

The European Commission goes beyond recommendations, and is drafting a binding Regulation for an Artificial Intelligence Act (European Commission, 2021) with the aim to ensure "the development, use and uptake of artificial intelligence in the internal market that at the same time meets a high level of protection of public interests, such as health and safety and the protection of fundamental rights" (page 19). This proposal takes after the General Data Protection Regulation (GDPR) approach of the EU – where strict regulations at first met opposition but turned into global inspiration (Craglia, 2018) – and embodies the EU's vision for "human-centric AI" (Digital Future Society, 2021). Note that the GDPR already contained a "right to explanation" (Doshi-Velez et al., 2017; Goodman and Flaxman, 2016. Although the precise wording of the AIA is likely to change before it's ratification, we assume there will be no significant changes to it's spirit and the draft regulation can be used to frame the regulatory context in which ML systems developed or deployed in Europe will need to adhere to.

The AIA Article 13 on transparency and prevision to users states: "high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately". High-risk systems are required to document the usage of the system which contains instructions including possible limitations and "the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users" (European Commission, 2021).

As to be expected, the emergence of these guidelines has led to many reactions from scholars involved in explainable AI methods. The main observation and consensus is that there is a gap between the explicability and transparency stipulated by the AIA and the capabilities of current technological explainable AI methods (Hamon et al., 2020; Sovrano et al., 2022). The use of ambiguous terms such as 'sufficiently' transparent and 'appropriate' types of transparency methods (Fink, 2021; Smuha et al., 2021) may lead to ethics washing. Yet the establishment of application-specific standards over the coming years may provide a stronger framework for conformity assessments before the AIA is expected to be adopted in 2024–2025 (Veale and Borgesius, 2021). For example the CEN-CENELEC is looking into "research-based metrics" and ISO/IEC TR 24028:2020(E) describes key metrics for explanations (Sovrano et al, 2022). Interestingly, it was observed that the draft AIA addresses the explainability of models but does not require the use of interpretable models (Sovrano et al., 2022). Finally, a common criticism of the AIA is that explainability is user-focused but does not consider explainability to the persons affected by the outputs of the algorithm (Fink, 2021; Smuha et al., 2021).

The definition of transparency in literature on explainable ML (see Section 2) generally considers the transparency of the ML model and it's parameters. Indeed, Roscher et al. (2020b) specifically state that the data itself is not considered in their conceptualization of transparency. Yet transparency has a broader definition in policy contexts, including a description of the input data and its distribution as well as how the output of the model actually leads to the decision (Hamon et al., 2020).

Sovrano et al., (2022) observe that the concept of explainability in AIA is *user-empowering* (i.e. empowering the user to interpret the system's output and use it appropriately) and *compliance oriented*. The ML system must therefore be sufficiently explainable to third-party auditors to demonstrate compliance (to-be-developed) standards before being released to the market. According to their review of the AIA, explainability metrics specifically for regulatory conformity assessments should: (i) be able to assess the risks to the fundamental rights of persons affected by the system's output, (ii) be model-agnostic and applicable to the wide range of methods falling under the AIA, (iii) flexible to the needs of the questioner, and (iv) intelligible and accessible.

The regulatory context, as evident from the UNESCO recommendations and the AIA, put a different nuance on explainable ML than the definitions provided above. Firstly, the regulatory context considers the

entire workflow of a decision-making train including data collection, the AI algorithm itself, and how the predictions are used for decision-making. This is a broader focus than only the algorithmic input and output as often considered in Earth Observation applications. Secondly, there are clearly different audiences for ML applications. Explanations are particularly important for model users to consider whether they can use a trained model for their application, auditors to consider whether a model meets the requirements of conformity assessments, and those affected to request an explanation on why a decision was taken. Especially for these applications, it is important that explanations are goal-aware and can be understood by users without expert domain knowledge.

5. Problems identified by social sciences

Miller (2019) makes a sharp critique of research on explainable machine learning by emphasizing that there is a gap between the techniques being developed and research from philosophy, psychology and cognitive science on human decision-making. Three main gaps are identified as human explanations tend to be: contrastive, selective by focusing on one or two reasonings for the outcome rather than all, and part of a social interaction to transfer knowledge. He also observes that stating causes of predictions is more useful for human understanding than providing statistical probabilities.

A second main critique on the field of explainable ML is that publication bias is driving research towards complex models with a high predictive accuracy while omitting simpler models with comparable predictive accuracy but lower descriptive accuracies (Rudin, 2018). Indeed, the published works on explainable ML in EO presented below indicates that the vast majority of works only present an explainability method without actually testing how useful the explanations are for the understanding of the system.

Thirdly, research tends to present innovative explainability methods without actually focusing on the utility of these new methods to foster explainability for real-world problems and users (Murdoch et al., 2019; J. Zhu et al., 2018). Researchers often present cherry-picked "reasonable" explanations, without rigorously testing the descriptive accuracy of the interpretation method (Murdoch et al., 2019). For example, published works tend to show saliency maps of the predicted class label and not of incorrect labels. This obfuscates the phenomenon that saliency maps for differing labels may highlight the same part of the image (Rudin, 2018) – a great limitation of using saliency maps for explanation. The limitations of saliency maps will be discussed in more detail below.

Some recent works attempt to quantify the descriptive accuracy of an explainable ML model, thereby enabling the trade-off between predictive accuracy and descriptive accuracy to be measured and providing some guidance to which is more appropriate. Three examples will be discussed in more detail. Firstly, desiderata of good explainability models can be defined. For example, Sovrano et al. (2022) refer to Carnap's central criteria of explanation accuracy (Leitgeb and Carus, 2021). Their first desiderata is *similarity*, which refers to how similar the explanation is to the workings of the model. This is similar to the concept of model fidelity described previously. Secondly, *exactness,* which refers to how clear the information is in terms of pertinence and syntax. Finally, *fruitfulness* refers to how useful a certain piece of information is to generate explanations. Note that none of the desiderata actually refer to the truthfulness of the explanation model.

Murdoch et al. (2019) describe a framework which consists of: predictive accuracy, descriptive accuracy, and relevance. The predictive accuracy refers to how accurate a model prediction is compared to reference data. I.e. the accuracy which is traditionally reported by ML algorithms. The descriptive accuracy refers to "the degree to which an interpretation method objectively captures the relationships learned by machine-learning models" and relevancy refers to explanations that "provide insight for a particular audience into a chosen domain problem" (Murdoch et al., 2019).

Rosenfeld (2021) presents a framework for quantifying the tradeoff between prediction accuracy and descriptive accuracy for different models. She presents four metrics, based on: (1) the difference between the predictive accuracy of the interpretable model and the predictive accuracy of the less interpretable model; (2) the number of rules in the explanation; (3) the number of features used to construct the explanation; and (4) the stability of the agent's explanation. She argues that in some cases, a very great increase in the predictive accuracy of a uninterpretable model over that obtainable through interpretable methods can make the loss of descriptive accuracy acceptable. This is somewhat at odds with the arguments by Rudin (2018) that uninterpretable models are unsuitable for high-stakes applications.

In conclusion, observations from the social sciences describe a number of key limitations of data-science driven explainability metrics up to now. Firstly, there is a discord between explainability models being developed and types of reasonings typically used by humans in explanations. Secondly, publications on ML topics tend to focus on new, complex methods with a high predictive accuracy without benchmarking them against simpler, more interpretable models. Thirdly, many new explainability methods are simply presented without showing the utility of their explanations through user testing. These limitations have inspired a number of frameworks to quantify the utility of explainability methods and potential trade-offs between predictive accuracy and descriptive accuracy.

6. Explainable ML methods in Remote Sensing and Earth Observation

Current trends in explainable ML for Remote Sensing and Earth Observation were analyzed by conducting a Scopus query: ("explainable AI" OR "XAI" OR "interpretability" OR "explainable" OR "interpretable" OR "explainability") AND ("remote sensing" OR "earth observation") AND ("artificial intelligence" OR "machine learning") and augmented through snowballing to obtain a list of 77 publications. Twenty-four papers were removed because they could not be located or they did not fit the scope of the intended literature analysis. Eight more consisted of high-level papers describing, e.g. the general need of explainability in Earth Observation applications. Of the remaining 45 works, only 8 were published before 2020. These 45 studies were analyzed to identify: the specific motivation of incorporating explainability into the ML workflow, the type of explanation method utilized, the intended audience of the explainability method, and whether or not the explainability method was actually evaluated.

6.1. *Why explainability*

Adadi and Berrada (2018) identify four underlying motivations for explainable ML: (1) explain to justify, (2) explain to control, (3) explain to improve, and (4) explain to discover. *Explain to justify* is set in the societal context of increased concern of the black-box nature of algorithms and addresses the need to investigate the reasonings behind the algorithms in order to justify why the model obtained a certain output prediction. This is closely linked with belief that an increased interpretability of ML models will lead to increased trust in their predictions (Doshi-Velez et al., 2017; Lipton, 2016; Miller, 2019) and the awareness that many ML applications demand accountability and therefore amenable to scrutiny (Camps-Valls et al., 2020; Doshi-Velez et al., 2017; Lipton, 2016; Roscher et al., 2020b; Tuia et al., 2021). Indeed a convincing argument for the use of ML methods to support decision-making is that algorithms are impartial to cognitive biases that plague humans and therefore using algorithmic explainability mechanisms may support more impartial decision-making (Arrieta et al., 2019; Doshi-Velez et al., 2017).

Explain to control implies that explainability can help identify possibly erroneous system behavior and speed up error debugging and

removal of flaws (Doshi-Velez et al., 2017; Lapuschkin et al., 2019). Simply understanding how models work is a cited motivation for explainability (Miller, 2019; Roscher et al., 2020a, 2020b). Similarly, *explain to improve* hypothesizes that an improved understanding of how models make predictions will accelerate the development of better models. For example models that are more robust (Arrieta et al., 2019; Doshi-Velez et al., 2017; Roscher et al., 2020a, 2020b) and have improved transferability (Lipton, 2016). Another common motivation for explainability along this vein is the belief that explainability can help develop more meaningful models (Lipton, 2016). In the context of Earth Observation this often means models that are loyal to the underlying physical principles that guide the natural processes that are being modelled (Camps-Valls et al., 2020; Roscher et al., 2020a, 2020b; Tuia et al., 2021).

Finally, *explain to discover* refers to the potential of explainable methods to help scientists discover new linkages and knowledge (Roscher et al., 2020a, 2020b). For example, identifying the features driving the model to a certain prediction can identify which features should be further analyzed for causal relationships (Lipton, 2016).

The reviewed publications on explainable ML in EO often report more than one of these four reasons to explain (Fig. 1). By far the most prominent motivation in these published works is *explain to control*. Often simply referred to as model interpretability, the authors describe how explainable ML methods can be used to find important factors and investigate whether the most relevant features identified by a model can be confirmed by domain knowledge. *Explain to improve* was relevant when, for example, the most important features were selected to train a new, sparser model while retaining a high accuracy. *Explain to discover* was the second most prominent motivation underlying the incorporation of explainable ML methods and *explain to justify* was only identified by three publications.

6.2. Explainability methods

Published methods to incorporate explainability into ML workflows in the domain of Earth Observation can be divided into four categories: interpretable models, incorporating domain knowledge, feature selection, and saliency maps (Fig. 2, Table 2). The first two categories consist of predominantly intrinsic, model-specific, global methods. Feature selection methods can be intrinsic or post-hoc depending on whether the selected features are actually utilized to change the model or the feature importance is only influenced in a post-hoc manner. Saliency maps are a type of explanation method specifically designed for CNNs and are generally post-hoc and local.

6.2.1. Interpretable models

Some researchers explicitly opt for the use of models that are more easily interpretable. This includes the use of Gaussian processes to estimate crop yield (Martinez-Ferrer et al., 2021; Mateo-Sanchis et al., 2021), the use of fuzzy logic to estimate the severity of disasters (Rodríguez et al., 2011), Generalized Linear Models to model dengue vector populations (Mudele et al., 2021), unsupervised clustering to forecast river sedimentation (N. Ahmed et al., 2019), Ordinary Differential Equations to estimate bioclimactic variables (Adsuara et al., 2020), and Logical Analysis of Data for hyperspectral image classification (A.M. Ahmed et al., 2019).

Other researchers aim at modifying 'black-box' methods such as CNNs to make them more interpretable. Dechesne et al., (2021) utilize Monte Carlo drop-out during the training phase to obtain a Bayesian deep learner a semantic segmentation task of identifying buildings in satellite imagery, thereby adding a level of confidence to the output prediction map. Stomberg et al., (2021) perform a clustering on intermediate steps of the CNN to identify the main 'concepts' that the network identifies as relating to the predicted variable of 'wilderness'. Levering et al., (2020) introduce semantic bottlenecks into the network architecture in order to classify images that are aesthetically pleasing, i.e. have a high 'scenicness'. Based on domain knowledge, they

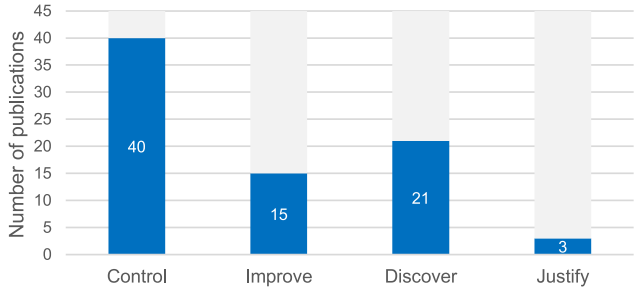


Fig. 1. Motivations underlying the integration of explainability into ML workflows for Earth Observation as perceived by published works.

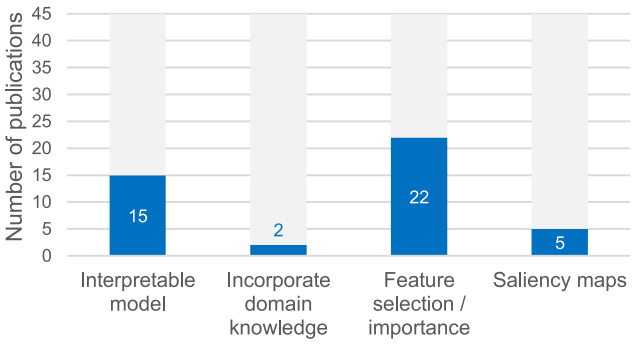


Fig. 2. Types of explanation methods in published works on explainable ML in Earth Observation.

Table 2
Categorization of explainable ML methods identified in Earth Observation publications according to the type of explanation method and the category it pertains to, where I = Intrinsic, H = Post-hoc, MS = Model-specific, AG = Model-agnostic, L = Local and G = Global. See Table 1 for more details on the different categories.

Type of explanation method	Category	N	References
Interpretable model	I-MS-G	14	(Adsuara et al., 2020; A. M. Ahmed et al., 2019; N. Ahmed et al., 2019; Dechesne et al., 2021; Dinc & Parra, 2021; Feng et al., 2021; Ghosh et al., 2020; Lacoste et al., 2011; Martinez-Ferrer et al., 2021; Mateo-Sanchis et al., 2021; Mudele et al., 2021; Rodríguez et al., 2011; Stomberg et al., 2021; Yan et al., 2021)
	I-MS-L	1	(Levering et al., 2020)
Incorporate domain knowledge	I-MS-G	2	(Kraft et al., 2020; Svendsen et al., 2021)
Feature selection	I-MS-G	3	(Mudele et al., 2020; Paudel et al., 2021; Stroppiana et al., 2021)
	H-MS-G	7	(Browne et al., 2021; Kirkwood et al., 2016; Murray et al., 2020, 2021; Newman & Furbank, 2021; Taconet et al., 2021; Upadhyaya et al., 2021)
	H-AG-G	5	(Duro et al., 2012; Fu et al., 2020; Guidici & Clark, 2017; Orynbaikyzy et al., 2020; Tian et al., 2021)
	H-AG-L/G	7	(Abdollahi & Pradhan, 2021; Chen et al., 2020; Ebrahimi-Khusfi et al., 2021; Han et al., 2022; Islam et al., 2020; Matin & Pradhan, 2021; Xing & Sieber, 2021)
Saliency maps	H-MS-L	5	(Huang et al., 2022; Hung et al., 2021; Kakogeorgiou & Karantzalos, 2021; Maddy & Boukabara, 2021; Wolanin et al., 2020)

hypothesize that the 'scenicness' score assigned to an image by volunteers will be related to the land-cover that is contained in that image. So they design a CNN that first classifies the land cover classes present in satellite imagery, and then performs a regression task to estimate the 'scenicness'. The landcover classes thereby form the semantic bottleneck that can be used to explain why the deep learning architecture identifies some scenes as being more aesthetic than others.

Utilizing ML models that are interpretable-by-design has a clear advantage for more explainable ML. The use of interpretable models can sometimes achieve higher (Mudele et al., 2021) or comparable (A.M. Ahmed et al., 2019) predictive accuracies than less interpretable models. However, intrinsic explanations methods are, by definition, model-specific and may be more difficult to evaluate than model-agnostic methods. Intrinsic explainability methods tend to be global. One notable exceptions in the list of reviewed works are the use of semantic bottlenecks (Levering et al., 2020), which provides local explanations.

Furthermore, there is no clear distinction between which models can be considered as explainable and which are not. For example, a decision tree is interpretable, but as decision trees grow or Random Forests are used, the relations become so complex that it can be questioned whether a human is still able to understand the decision-making process (Murdoch et al., 2019). This questions interpretability assumptions in the domain of EO, where Random Forests are commonly considered to be more interpretable than deep learning or SVM alternatives. For example, Browne et al., (2021) use Random Forests with 2000 trees to predict poverty and malnutrition using openly available spatial data; Kirkwood et al., (2016) utilized a Regression orest with 1001 trees for geochemical mapping; and Newman and Furbank, (2021) combine thousands of variables in Random Forests to predict crop traits. All of these studies assert the usage of Random Forests as interpretable models, though the complexity of the resulting model limits the actual interpretability to understanding the global importance of input variables.

6.2.2. Incorporating domain knowledge

Another group of intrinsic explanation methods specifically aim to incorporate domain knowledge into the machine learning workflow. These hybrid methods aim to strike a balance between the high explainability but often low accuracy of traditional frameworks based on expert knowledge vs. the high accuracy but low explainability of data-driven machine learning methods (Svendsen et al., 2021). Kraft et al., (2020) provide a hybrid modeling framework to global hydrological modelling. It combines a neural network with a water balance model which adds restraints based on the natural processes such as evapotranspiration and run-off. Svendsen et al., 2021 provide another hybrid model which utilizes latent force models to incorporate natural relations in soil moisture and biophysical parameters over time series.

Incorporating domain knowledge into ML using the strategies described here result in intrinsic, global, model-specific methods. Constraining ML algorithms by the restraints defined by natural processes will likely result in more reliable ML algorithms. The importance of multi-disciplinary teams in the design of these models is key to ensure their veracity as well as to investigate whether patterns learned by the ML models can further domain knowledge.

6.2.3. Feature selection and importance

Feature selection can be used to induce sparsity and enhance the explainability of models (Murdoch et al., 2019). Similarly, ranking the importance of features can identify the importance of underlying physical processes (e.g. Martinez-Ferrer et al., 2021; Mateo-Sanchis et al., 2021) or which features were key to the model outcome for one specific sample (e.g. Matin and Pradhan, 2021). Feature selection was the most common type of explainable ML method identified in the literature review, representing 22 out of 45 publications. For the purposes of the current review, feature selection methods were considered as post-hoc when used to rank feature importance after the modelling of the classifier (19 out of 22). If, however, feature selection was conducted before training a model in order to enforce sparsity, then it was labelled as an intrinsic part of the ML workflow (3 out of 22). Feature selection for machine learning techniques is well established in the field of Earth Observation (Belgiu and Drăguţ, 2016; Bruzzone and Serpico, 2000; Camps-Valls, 2009), though perhaps not connected to the terms "explainability" and "interpretability" until more recently. This observation is supported by the recency of the articles on explainable ML and EO reviewed in this work (e.g. Fu et al., 2020; Orynbaikyzy et al., 2020).

However, the appearance of Shapely Additive exPlanations (SHAP) (Lundberg et al., 2017) in the domain of Earth Observation is relatively new. Seven of the reviewed publications use SHAP. This method was developed in the computer science community and utilizes game theory to identify relevant features in a model-agnostic manner. It can be used to identify important features for the model in general (i.e. global explanations) or for a specific prediction (i.e. local explanations) (Molnar, 2022). Publications in the field of EO have shown that SHAP can be used to gain insights to urban vegetation mapping (Abdollahi and Pradhan, 2021), land cover classification (Xing and Sieber, 2021), understand factors driving housing prices (Chen et al., 2020), determine environmental variables driving dust pollution (Ebrahimi-Khusfi et al., 2021), and assessing building damage after earthquakes (Matin and Pradhan, 2021). These developments should encourage researchers in the Earth Observation domain to keep an eye on emerging explainability techniques from computer science, though more research is also needed to understand whether and how these feature selection techniques compare with those already integrated in the domain of Earth Observation.

6.2.4. Pixel attribution maps

Pixel attribution maps, also known as saliency maps or heat maps, is a family of explanation maps specifically designed for convolutional neural networks. These maps visualize the influence of different sections of the image on the model's prediction for that image (Molnar, 2022). They can be used by model developers to find obvious errors in the model, for example Ribeiro et al., (2016) describe how a CNN-classifier trained to discriminate huskies from wolves actually focused on the snow in the background of the images rather than the facial characteristics of the animals themselves. Pixel attribution maps have gained much attention from the domain of computer vision as a possible way to visualize the workings of 'black-box' deep CNNs and some methods specifically tailored for earth observation imagery have recently been proposed (Huang et al., 2022; Hung et al., 2021).

Kakogeorgiou and Karantzalos (2021) conducted a comprehensive study comparing saliency map methods for remote sensing applications. The authors compared the results of 12 documented saliency map methods on two RS benchmark datasets for the classification of satellite imagery. The methods were compared using five metrics: the sensitivity of the output map to small input perturbations, how quickly prediction decreases as 'salient' pixels are removed, the file size (i.e. complexity) of the output map, and computation time. The best interpretability was reported for Occlusion (Zeiler and Fergus, 2014), Grad-CAM (Selvaraju et al., 2017), and Lime (Ribeiro et al., 2016), whereas Guided Backpropagation (Springenberg et al., 2014) had the lowest reliability.

However, these studies do not consider the limitations of using saliency maps in general to provide explanations of black-box models. Trained models have been shown to produce similar saliency maps to randomly initialized models (Adebayo et al., 2018) and backpropagation-based visualizations have been shown to perform image recovery rather than providing information on network decisions (Nie et al., 2018). Saliency maps for different classes may also highlight the same part of the image (Rudin, 2018), thereby limiting their explanatory value. That is to say, if a saliency map for labelling an image as "husky" and a saliency map for labelling the same image as "flute" highlight the same area of the image then what is the added value of a saliency map?

Stomberg et al., (2021) developed a method to actually integrate

pixel attribution maps into a deep learning workflow for classifying wilderness areas from satellite imagery. A classification task is set-up to identify 'wilderness' areas, but as the concept of wilderness in satellite imagery is difficult to describe, the activation maps of the bottleneck of a U-NET architecture are clustered to identify 'concepts' that distinguish wilderness from non-wilderness areas. They thereby transform a typical post-hoc explainability method into an intrinsic explainability method with the underlying motivation of generating new knowledge.

6.3. Intended audience & validation of the explainability method

Finally, the list of publications on explainable ML in EO were analyzed to find the intended audience of the explanation and whether the provided explanations were tested in practice. The audience for the explanations provided by almost all publications were experts and researchers. The exceptions are (Murray et al., 2020), who provided linguistic explanations comparing land cover class performance of different deep learning models for non-fusion experts; and Rodríguez et al., (2011) who utilized fuzzy rules and linguistic labels to help decision-makers understand the severity of a natural disaster.

Similarly, most publications simply presented the results of the explainable ML method without testing the usefulness of these explanations in practice. Kakogeorgiou and Karantzalos, (2021) rigorously compare the performance of different saliency map algorithms and Wolanin et al., (2020) compare regression activation maps to physical variables. Yet no study assessed to which degree the use of explainability methods actually helps users understand the algorithm or how such explanations can speed up and improve the development of ML workflows.

7. Discussion

7.1. Explainable ML in Earth Observation and the regulatory context

The regulatory context considers transparency in a broader context than that considered in the EO publications reviewed above. This implies that more work is needed to understand how to document input data and help users understand the applicability of the model for a specific application. Hamon et al. (2020) points out that this can be done through the development of fixed dataset descriptions (Gebru et al., 2021) and model cards (Mitchell et al., 2018). However these would still need to be developed in the domain of EO.

Secondly, the regulatory context implies that explanations should be goal-aware. Specific goals include conformity assessments, users of algorithmic workflows, and appeals to specific decisions. Yet the review of EO literature above indicated that only two out of the 45 publications utilized an explanation that was developed for someone without specific domain knowledge. This is a clear gap for ML explanations that are suitable for non-experts. Furthermore, there is a particular consideration for conformity assessments. Sovrano et al. (2022) points out that explainability methods for conformity assessments should be model-agnostic. However, it has been shown that common explainable ML methods from Computer Vision need to be adapted to the unique characteristics of EO data (e.g. Camps-Valls et al., 2021; Xing and Sieber, 2021). Once the standards are developed, research should therefore be conducted to ensure that these model-agnostic methods are also suitable for EO data.

Recognizing that EO data may require different workflows, main players in the geospatial industry are calling for companies to contribute to the development of best-practices which could be used as examples for the developments of standards (WGIC, 2021). Although it is applaudable that companies personally take responsibility for the ethical usage and development of ML systems, it is important that these best-practices are subject to wider debate to prevent best-practices from turning into cherry-picked shortcuts (e.g. Borsci et al., 2022).

7.2. Explainable ML in Earth Observation and concerns from the social sciences

Section 5 identified three shortcomings of explainable ML methods that were identified by social sciences. Firstly, that explainability mechanisms developed by explainable ML differ from the types of explanations that humans use tend to focus on, i.e. contrastive explanations and the selection of a few key examples to explain. Yet the explainability methods presented in the Earth Observation domain are much more complex and depend on many different features. This is not surprising as Earth Observation is similar to Computer Vision, which is typically a difficult domain for interpretability (Rudin, 2018). Still, the integration of explainable ML methods that imitate human decision making should be investigated. For example, counterfactuals, which aim to describe the minimal change to the input that would result in a different prediction is one such example (Rudin, 2018) and case-based explanations which look for similar examples in the training data to justify the recommendations (Nugent et al., 2009).

The second shortcoming was the publication bias which focusses on complex models with a high predictive accuracy without benchmarking them against simpler, more interpretable methods. The review of Earth Observation literature indicated that 34% of the publications benchmarked a less explainable model against a simpler, more explainable model. However, this could also mean comparing a deep CNN against Random Forests or SVM classifiers. Although these publications describe Random Forests as an interpretable method, they are generally not accepted as interpretable by the social sciences as their workings are too complex for humans to easily understand (Murdoch et al., 2019). Assessment frameworks that specifically assess the predictive accuracy and descriptive accuracy together can help quantify the trade-off and define which criteria a model should fit in order to be considered interpretable (e.g. Murdoch et al., 2019; Rosenfeld, 2021).

The third shortcoming of explainable ML was the lack of consideration of whether the explainable methods actually work in context. This issue extends to the domain of Earth Observation as almost all the explainability methods in the reviewed works targeted domain experts and model developers; and there is a tendency to simply present "explanations" without assessing whether these explanations are actually helpful for the intended audience.

7.3. Which type of explanation do we want?

Given the known fallacies of explainable ML methods and upcoming regulatory frameworks, researchers in the EO field should carefully consider which types of explanations they intend to use. It is recommended to consider of the intended purpose and audience of the explanation when selecting which explainability method to use. Fig. 3 provides an example flow chart to support this selection.

First, verify that whether ML methods with intrinsic explainability have the same predictive accuracy as 'black-box' methods. Use simpler models with intrinsic explainability if they achieve similar accuracies. Note, that contrary to the statements of many publications in the domain of Earth Observation, very complex Random Forests should not be considered explainable. Frameworks such as Rosenfeld (2021) and Murdoch et al., (2019) can help quantify and guide this trade-off between predictive and descriptive accuracy.

However, if complex models have a significantly higher predictive accuracy, then incorporate a post-hoc explainability method into the ML workflow. The type of post-hoc explainability method will depend on the intended purpose and audience of the explanation. *Researchers* interested in obtaining new domain knowledge will be interested in the model-agnostic relationship between natural phenomena in the real world rather than the outcomes of one specific model. *Model developers* will use both global explanations to understand the general workings of the model as well as local explanations to debug it. For example, the use of saliency maps can help check for obvious errors in the model, but they

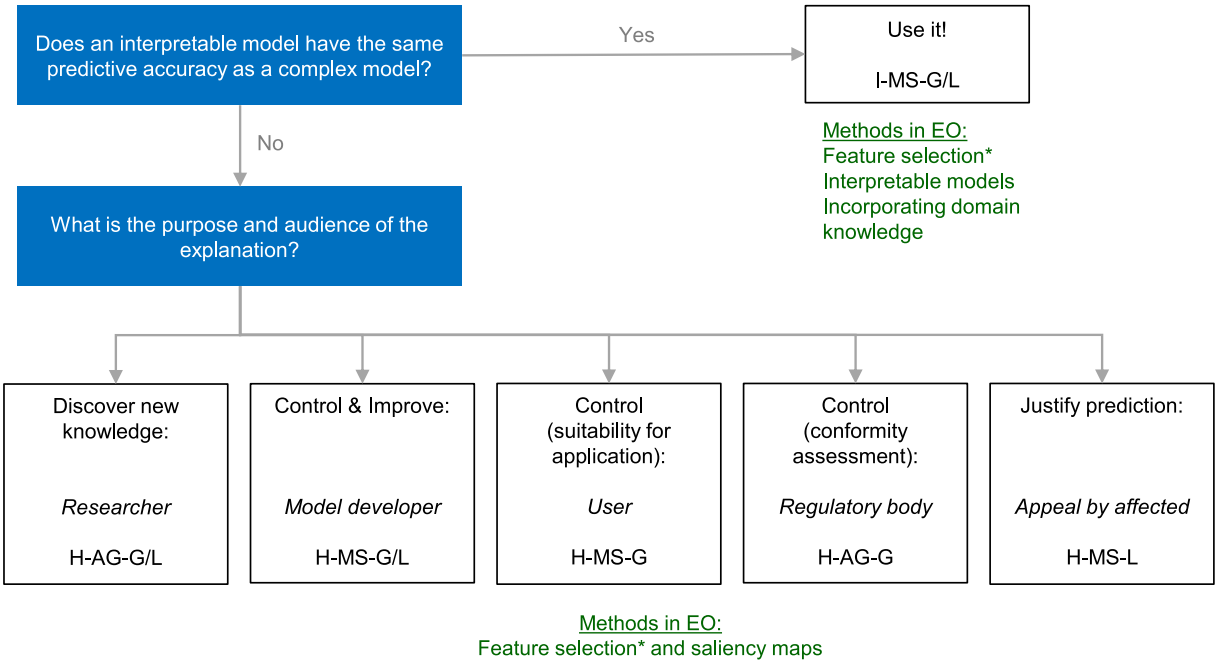


Fig. 3. Flowchart to help select a type of explainability method based on the intended user. The recommended methods can be intrinsic or post-Hoc (I/H), model specific or model agnostic (MS/AG), and global or local (G/L). In green, explainability methods appearing in published works in the domain of EO are listed. See Table 2 for more details on the EO methods and the text for discussions regarding the limitations of these methods for the various explanation purposes and audiences.

will not help to fully understand how a prediction was made. *Users* of a model developed by a third party will need to have a sufficient understanding of the general strengths and weaknesses of a model in order to judge whether they are able to confidently deploy the model in their specific context. *Regulatory bodies* developing conformity assessments, such as the EU, will require model-agnostic, global methods (Sovrano et al., 2022). The users of ML models developed by a third party will need to have a sufficient understanding of the global strengths and weaknesses of a specific model to ensure that they use the model correctly. Finally, in situations where *someone affected by the output* of an algorithm would like to appeal to the prediction and understand the model behavior that led to a the prediction outcome of that particular case (i.e. explain to justify), then a model-specific method with a local explanation should be selected. Note that Fig. 3 is of course a general overview of common cases, and specific preferences will depend heavily on the specific application and user.

An alarming gap is revealed when comparing the types of explanations required in Fig. 3 with the list of methods utilized in EO in Table 2. For example, Fig. 3 indicates that local, model-specific explanations are required in order to justify the result of a ML model to a person affected by it. Yet the only post-hoc model-specific methods utilized in EO are saliency maps, which, as discussed above, are not easily interpretable. Similarly, the only other type of post-hoc method identified in Table 2 is feature-selection. So at the moment, feature selection and saliency maps are the only published types of post-hoc explanations in the EO community which could be applied to the use cases in the bottom row of Fig. 3 (researchers, model developers, users, regulatory bodies, and appeals). As the domain of EO starts adopting explainability techniques, it is important to consider the purpose and audience of the explanation and critically select which type of explanation would be suitable.

Similarly, almost all of the reviewed publications focused on developing explanations for expert users. Yet remote sensing is increasingly being used to distribute aid for humanitarian (Lang et al., 2020) or disaster risk reduction (Deparday et al., 2019) purposes in LMICs. For these applications, it becomes paramount to develop explainability mechanisms that can justify why some households should receive more aid than others and communicate the uncertainties of these reasonings to humanitarian actors.

8. Conclusions and recommendations

Although understanding ML models has been part of domain of Earth Observation for some time, the new wave of explainable ML driven by societal implications and regulatory frameworks is just starting. Rather than blindly copying explainability methods developed in the domain of computer science, experts in the domain of Earth Observation should be critical of the weaknesses of these methods and aim to correct them rather than copy them. This insinuates critically assessing the usefulness of explanations produced by explainability methods and whether they are appropriate for the intended audience and application.

This review of existing methods for explainability in ML resulted in a series of recommendations on how to select a relevant explainability method depending on the context of the issue. In case of similar predictive accuracies, always use the simpler, interpretable model rather than an uninterpretable comparison (Murdoch et al., 2019; Rudin, 2018). Use complex networks when the accuracy achieved through complexity trumps the explainability, and does not fundamental risk for infringement of human rights. If complex networks are utilized, safeguard explainability by clearly describing and specifications of the training dataset, training procedure, and accuracies (Hamon et al., 2020) and choosing a suitable goal-aware explainability mechanism. The overview in Fig. 3 can support the latter.

In particular, the domain of Earth Observation shows limitations regarding: which algorithms are considered interpretable; the availability of post-hoc methods suitable for different purposes and audiences of explanations; the development of explainability methods for non-experts; and the lack of rigorous testing of the quality of the produced explanations. As we move forward to tackle these challenges, keeping an eye on other domains such as social sciences and emerging regulatory requirements can help ensure that the methods we develop are suitable for a broader range of stakeholders. And that thus earth observation and machine learning can be more responsibly deployed to tackle the

humanitarian and sustainability issues of our time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This publication is part of the project "Bridging the gap between artificial intelligence and society: Developing responsible and viable solutions for geospatial data" (with project no. 18091) of the research program Talent Program Veni 2020, which is (partly) financed by the Dutch Research Council (NWO).

The author would especially like to thank Prof. Yola Georgiadou and Dr. Andrea Aler Tubella for their support and feedback regarding the conceptualization of this work.

References

Abdollahi, A., Pradhan, B., 2021. Urban vegetation mapping from aerial imagery using explainable AI (XAI). Sensors 21 (14), 4738. https://doi.org/10.3390/s21144738.

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity Checks for Saliency Maps.

Adsuara, J.E., Perez-Suay, A., Moreno-Martinez, A., Camps-Valls, G., Kraemer, G., Reichstein, M., Mahecha, M., 2020. Discovering Differential Equations from Earth Observation Data. Int. Geosci. Remote Sensing Symposium (IGARSS) 3999–4002. https://doi.org/10.1109/IGARSS39084.2020.9324639.

Ahmed, A. M., Ibrahim, S.K., Yacout, S., 2019. Hyperspectral Image Classification Based on Logical Analysis of Data. IEEE Aerospace Conference Proceedings, 2019-March. https://doi.org/10.1109/AERO.2019.8742023.

Ahmed, N., Mahmud, S., Lutfe Elahi, M.M., Ahmed, S., Sujauddin, M., 2019b. Forecasting river sediment deposition through satellite image driven unsupervised machine learning techniques. Remote Sens. Appl.: Soc. Environ. 13, 435–444. https://doi.org/10.1016/j.rsase.2018.12.011.

Arrieta, A. B., Díaz-Rodríguez, N., del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. http://arxiv.org/abs/1910.10045.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/J.ISPRSJPRS.2016.01.011.

Biran, O., Cotton, C., 2017. Explanation and justification in machine learning: a survey. IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI).

Borsci, S., Lehtola, V. v., Nex, F., Yang, M. Y., Augustijn, E.-W., Bagheriye, L., Brune, C., Kounadi, O., Li, J., Moreira, J., van der Nagel, J., Veldkamp, B., Le, D. v., Wang, M., Wijnhoven, F., Wolterink, J. M., & Zurita-Milla, R. (2022). Embedding artificial intelligence in society: looking beyond the EU AI master plan using the culture cycle. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01383-x.

Browne, C., Matteson, D.S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., Barrett, C.B., Forkuor, G., 2021. Multivariate random forest prediction of poverty and malnutrition prevalence. PLoS ONE 16 (9), e0255519. https://doi.org/10.1371/journal.pone.0255519.

Bruzzone, L., Serpico, S.B., 2000. A technique for feature selection in multiclass problems. Int. J. Remote Sens. 21 (3), 549–563. https://doi.org/10.1080/014311600210740.

Burrell, J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data and Society, 3(1). https://doi.org/10.1177/2053951715622512.

Camps-Valls, G., 2009. Machine learning in remote sensing data processing. IEEE International Workshop on Machine Learning for Signal Processing 2009, 1–6. https://doi.org/10.1109/MLSP.2009.5306233.

Camps-Valls, G., Reichstein, M., Zhu, X., Tuia, D., 2020. Advancing Deep Learning for Earth Sciences: From Hybrid Modeling to Interpretability. International Geoscience and Remote Sensing Symposium (IGARSS) 3979–3982. https://doi.org/10.1109/IGARSS39084.2020.9323558.

Camps-Valls, G., Tuia, D., Zhu, X.X., Reichstein, M., 2021. Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences. In: Deep Learning for the Earth Sciences (Vol. 1). Wiley. https://doi.org/10.1002/9781119646181.fmatter.

Chen, L., Yao, X., Liu, Y., Zhu, Y., Chen, W., Zhao, X., Chi, T., 2020. Measuring impacts of urban environmental elements on housing prices based on multisource data-a case study of Shanghai, China. ISPRS Int. J. Geo-Inf. 9 (2), 106. https://doi.org/10.3390/ijgi9020106.

Craglia, M., 2018. Artificial intelligence : a European perspective. Publications Office. https://doi.org/10.2760/936974.

Dechesne, C., Lassalle, P., Lefèvre, S., 2021. Bayesian u-net: Estimating uncertainty in semantic segmentation of earth observation images. Remote Sensing 13 (19), 1–31. https://doi.org/10.3390/rs13193836.

Deparday, V., Gevaert, C.M., Molinario, G., Soden, R., Balog-Way, S., 2019. Machine Learning for Disaster Risk Management.

Digital Future Society, 2021. Governing algorithms: perils and powers of AI in the public sector.

Dinc, S., Parra, L.A.C., 2021. A three layer spatial-spectral hyperspectral image classification model using guided median filters. In: Proceedings of the 2021 ACMSE Conference - ACMSE 2021: The Annual ACM Southeast Conference, pp. 122–129. https://doi.org/10.1145/3409334.3452045.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., Wood, A., 2017. Accountability of AI Under the Law: The Role of Explanation. http://arxiv.org/abs/1711.01134.

Du, M., Liu, N., Hu, X., 2018. Techniques for Interpretable Machine Learning. http://arxiv.org/abs/1808.00033.

Duro, D.C., Franklin, S.E., Dubé, M.G., 2012. Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests. Int. J. Remote Sens. 33 (14), 4502–4526. https://doi.org/10.1080/01431161.2011.649864.

Ebrahimi-Khusfi, Z., Taghizadeh-Mehrjardi, R., Roustaei, F., Ebrahimi-Khusfi, M., Mosavi, A.H., Heung, B., Soleimani-Sardo, M., Scholten, T., 2021. Determining the contribution of environmental factors in controlling dust pollution during cold and warm months of western Iran using different data mining algorithms and game theory. Ecol. Ind. 132, 108287. https://doi.org/10.1016/j.ecolind.2021.108287.

European Commission Directorate-General for Communications Networks, C., and T., 2021. Proposal for a Regulation of the European Parliament and of the council Laying Down Harmonised Rules on ARtificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

General Data Protection Regulation (GDPR), Pub. L. No. Directive 95/46/EC, Regulation 2016/679, 2016.

Feng, S., Ji, K., Zhang, L., Ma, X., Kuang, G., 2021. SAR Target Classification Based on Integration of ASC Parts Model and Deep Learning Algorithm. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 10213–10225. https://doi.org/10.1109/JSTARS.2021.3116979.

Fink, M. (2021). *The EU AI Act and Access to Justice*. www.eulawlive.com.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M., 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.3518482.

Freitas, A.A., 2014. Comprehensible classification models. ACM SIGKDD Explorations Newsletter 15 (1), 1–10. https://doi.org/10.1145/2594473.2594475.

Fu, Z., Hu, L., Chen, Z., Zhang, F., Shi, Z., Hu, B., Du, Z., Liu, R., 2020. Estimating spatial and temporal variation in ocean surface pCO<inf>2</inf> in the Gulf of Mexico using remote sensing and machine learning techniques. Sci. Total Environ. 745 https://doi.org/10.1016/j.scitotenv.2020.140965.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K., 2021. Datasheets for datasets Vol. 64(12, 86–92. https://doi.org/10.1145/3458723.

Ghosh, S.M., Behera, M.D., Paramanik, S., 2020. Canopy height estimation using sentinel series images through machine learning models in a Mangrove Forest. Remote Sensing 12 (9), 1519. https://doi.org/10.3390/rs12091519.

Goodman, B., Flaxman, S., 2016. European Union regulations on algorithmic decision-making and a "right to explanation." https://doi.org/10.1609/aimag.v38i3.2741.

Guidici, D., Clark, M., 2017. One-dimensional convolutional neural network land-cover classification of multi-seasonal hyperspectral imagery in the San Francisco Bay Area, California. Remote Sensing 9 (6), 629. https://doi.org/10.3390/rs9060629.

Hamon, Ronan., Junklewitz, Henrik., Sanchez, Ignacio., & European Commission. Joint Research Centre, 2020. Robustness and explainability of Artificial Intelligence : from technical to policy solutions.

Han, L., Yang, G., Yang, X., Song, X., Xu, B.o., Li, Z., Wu, J., Yang, H., Wu, J., 2022. An explainable XGBoost model improved by SMOTE-ENN technique for maize lodging detection based on multi-source unmanned aerial vehicle images. Comput. Electron. Agric. 194, 106804. https://doi.org/10.1016/j.compag.2022.106804.

Herman, B., 2017. The Promise and Peril of Human Evaluation for Model Interpretability. http://arxiv.org/abs/1711.07414.

Holte, R.C., 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning 11 (1), 63–90. https://doi.org/10.1023/A:1022631118932.

Huang, X.u., Sun, Y., Feng, S., Ye, Y., Li, X., 2022. Better Visual Interpretation for Remote Sensing Scene Classification. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/LGRS.2021.3132920.

Hung, S.-C., Wu, H.-C., Tseng, M.-H., 2021. Integrating image quality enhancement methods and deep learning techniques for remote sensing scene classification. Applied Sciences (Switzerland) 11 (24), 11659. https://doi.org/10.3390/app112411659.

Islam, M.A., Anderson, D.T., Pinar, A.J., Havens, T.C., Scott, G., Keller, J.M., 2020. Enabling Explainable Fusion in Deep Learning with Fuzzy Integral Neural Networks. IEEE Trans. Fuzzy Syst. 28 (7), 1291–1300. https://doi.org/10.1109/TFUZZ.2019.2917124.

Kakogeorgiou, I., Karantzalos, K., 2021. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. Int. J. Appl. Earth Obs. Geoinf. 103, 102520. https://doi.org/10.1016/j.jag.2021.102520.

Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016. A machine learning approach to geochemical mapping. J. Geochem. Explor. 167, 49–61. https://doi.org/10.1016/j.gexplo.2016.05.003.

Kraft, B., Jung, M., Körner, M., Reichstein, M., 2020. Hybrid Modeling: Fusion of a Deep Learning Approach and a Physics-Based Model for Global Hydrological Modeling. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives 43 (B2), 1537–1544. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020.

Lacoste, M., Lemercier, B., Walter, C., 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133 (1–2), 90–99. https://doi.org/10.1016/j.geomorph.2011.06.026.

Lang, S., Füreder, P., Riedler, B., Wendt, L., Braun, A., Tiede, D., Schoepfer, E., Zeil, P., Spröhnle, K., Kulessa, K., Rogenhofer, E., Bäuerl, M., Öze, A., Schwendemann, G., Hochschild, V., 2020. Earth observation tools and services to increase the effectiveness of humanitarian assistance. European Journal of Remote Sensing 53 (sup2), 67–85. https://doi.org/10.1080/22797254.2019.1684208.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. 10 (1), 1096. https://doi.org/10.1038/s41467-019-08987-4.

Leitgeb, H., Carus, A., 2021. Rudolf Carnap. In: Zalta, E.N. (Ed.), *The Stanford Encoclypedia of Philosophy* (Summer 2021 Edition). Stanford University, Metaphysics Research Lab.

Levering, A., Marcos, D., Lobry, S., Tuia, D., 2020. Interpretable Scenicness from Sentinel-2 Imagery. International Geoscience and Remote Sensing Symposium (IGARSS) 3983–3986. https://doi.org/10.1109/IGARSS39084.2020.9323706.

Lipton, Z.C., 2016. The Mythos of Model Interpretability. http://arxiv.org/abs/1606.03490.

Lundberg, S.M., Allen, P.G., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), 1–10. https://github.com/slundberg/shap.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Maddy, E.S., Boukabara, S.A., 2021. MIIDAPS-AI: An Explainable Machine-Learning Algorithm for Infrared and Microwave Remote Sensing and Data Assimilation Preprocessing - Application to LEO and GEO Sensors. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 8566–8576. https://doi.org/10.1109/JSTARS.2021.3104389.

Martinez-Ferrer, L., Piles, M., Camps-Valls, G., 2021. Crop Yield Estimation and Interpretability with Gaussian Processes. IEEE Geosci. Remote Sens. Lett. 18 (12), 2043–2047. https://doi.org/10.1109/LGRS.2020.3016140.

Mateo-Sanchis, A., Piles, M., Amorós-López, J., Muñoz-Marí, J., Adsuara, J.E., Moreno-Martínez, Á., Camps-Valls, G., 2021. Learning main drivers of crop progress and failure in Europe with interpretable machine learning. Int. J. Appl. Earth Obs. Geoinf. 104, 102574. https://doi.org/10.1016/j.jag.2021.102574.

Matin, S.S., Pradhan, B., 2021. Earthquake-induced building-damage mapping using explainable ai (Xai). Sensors 21 (13), 4489. https://doi.org/10.3390/s21134489.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T., 2018. Model Cards for Model Reporting. https://doi.org/10.1145/3287560.3287596.

Molnar, C., 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

Mudele, O., Bayer, F.M., Zanandrez, L.F.R., Eiras, A.E., Gamba, P., 2020. Modeling the Temporal Population Distribution of Ae. Mosquito Using Big Earth Observation Data. IEEE Access 8, 14182–14194. https://doi.org/10.1109/ACCESS.2020.2966080.

Mudele, O., Frery, A.C., Zanandrez, L.F.R., Eiras, A.E., Gamba, P., 2021. Modeling dengue vector population with earth observation data and a generalized linear model. Acta Trop. 215, 105809. https://doi.org/10.1016/j.actatropica.2020.105809.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. PNAS 116 (44), 22071–22080. https://doi.org/10.1073/pnas.1900654116.

Murray, B.J., Anderson, D.T., Havens, T.C., Wilkin, T., Wilbik, A., 2020. Information Fusion-2-Text: Explainable Aggregation via Linguistic Protoforms. In Communications in Computer and Information Science: Vol. 1239 CCIS. https://doi.org/10.1007/978-3-030-50153-2_9.

Murray, B.J., Islam, M.A., Pinar, A.J., Anderson, D.T., Scott, G.J., Havens, T.C., Keller, J. M., 2021. Explainable AI for the Choquet Integral. IEEE Transactions on Emerging Topics in Computational Intelligence 5 (4), 520–529. https://doi.org/10.1109/TETCI.2020.3005682.

Newman, S.J., Furbank, R.T., 2021. Explainable machine learning models of major crop traits from satellite monitored continent-wide field trial data. Nat. Plants 7, 1354–1363. https://doi.org/10.1038/s41477-021-01001-0.

Nie, W., Zhang, Y., Patel, A., 2018. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations.

Nugent, C., Doyle, D., Cunningham, P., 2009. Gaining insight through case-based explanation. J. Intell. Inform. Syst. 32 (3), 267–295. https://doi.org/10.1007/s10844-008-0069-0.

Orynbaikyzy, A., Gessner, U., Mack, B., Conrad, C., 2020. Crop type classification using fusion of sentinel-1 and sentinel-2 data: Assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies. Remote Sensing 12 (17), 2779. https://doi.org/10.3390/rs12172779.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016. https://doi.org/10.1016/j.agsy.2020.103016.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " 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, pp. 1135–1144.

Rodríguez, J.T., Vitoriano, B., Montero, J., Kecman, V., 2011. A disaster-severity assessment DSS comparative analysis. OR Spectrum 33 (3), 451–479. https://doi.org/10.1007/s00291-011-0252-5.

Roscher, R., Bohn, B., Duarte, M.F., Garcke, J., 2020a. Explain it to me-facing remote sensing challenges in the bio-and geosciences with explainable machine learning. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 5 (3), 817–824. https://doi.org/10.5194/isprs-Annals-V-3-2020-817-2020.

Roscher, R., Bohn, B., Duarte, M.F., Garcke, J., 2020b. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 8, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.

Rosenfeld, A., 2021. Better Metrics for Evaluating Explainable Artificial Intelligence. AAMAS 45–50.

Rudin, C., 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. http://arxiv.org/abs/1811.10154.

Schorr, C., Goodarzi, P., Chen, F., Dahmen, T., 2021. Neuroscope: An explainable ai toolbox for semantic segmentation and image classification of convolutional neural nets. Appl. Sci. (Switzerland) 11 (5), 1–16. https://doi.org/10.3390/app11052199.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision (ICCV) 2017, 618–626. https://doi.org/10.1109/ICCV.2017.74.

Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., Maclaren, J., Piselli, R., Yeung, K., 2021. HOW THE EU CAN ACHIEVE LEGALLY TRUSTWORTHY AI: A RESPONSE TO THE EUROPEAN COMMISSION'S PROPOSAL FOR AN ARTIFICIAL INTELLIGENCE ACT.

Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F., 2022. Metrics, Explainability and the European AI Act Proposal. J 5 (1), 126–138. https://doi.org/10.3390/j5010010.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for Simplicity: The All Convolutional Net.

Stomberg, T., Weber, I., Schmitt, M., Roscher, R., 2021. Jungle-net: Using explainable machine learning to gain new insights into the appearance of wilderness in satellite imagery. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 5 (3), 317–324. https://doi.org/10.5194/isprs-annals-V-3-2021-317-2021.

Stroppiana, D., Bordogna, G., Sali, M., Boschetti, M., Sona, G., Brivio, P.A., 2021. A fully automatic, interpretable and adaptive machine learning approach to map burned area from remote sensing. ISPRS Int. J. Geo-Inf. 10 (8), 546. https://doi.org/10.3390/ijgi10080546.

Svendsen, D.H., Piles, M., Munoz-Mari, J., Luengo, D., Martino, L., Camps-Valls, G., 2021. Integrating Domain Knowledge in Data-Driven Earth Observation With Process Convolutions. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2021.3059550.

Taconet, P., Porciani, A., Soma, D.D., Mouline, K., Simard, F., Koffi, A.A., Pennetier, C., Dabiré, R.K., Mangeas, M., Moiroux, N., 2021. Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso. Parasites Vectors 14 (1). https://doi.org/10.1186/s13071-021-04851-x.

Tian, H., Wang, P., Tansey, K., Han, D., Zhang, J., Zhang, S., Li, H., 2021. A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China. Int. J. Appl. Earth Obs. Geoinf. 102, 102375. https://doi.org/10.1016/j.jag.2021.102375.

Tuia, D., Roscher, R., Wegner, J.D., Jacobs, N., Zhu, X., Camps-Valls, G., 2021. Toward a Collective Agenda on AI for Earth Science Data Analysis. IEEE Geosci. Remote Sens. Mag. 9 (2), 88–104. https://doi.org/10.1109/MGRS.2020.3043504.

UNESCO, 2021. Recommendation on the Ethics of Artificial Intelligence. In *UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000380455.

Upadhyaya, S.A., Kirstetter, P.E., Kuligowski, R.J., Searls, M., 2021. Classifying precipitation from GEO satellite observations: Diagnostic model. Q. J. R. Meteorolog. Soc. 147 (739), 3318–3334. https://doi.org/10.1002/qj.4130.

Veale, M., Borgesius, F.Z., 2021. Demystifying the Draft EU Artificial Intelligence Act. Computer Law Review International. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-.

WGIC, 2021. Geospatial AI/ML Applications and Policies: A Global Perspective. https://wgicouncil.org/wp-content/uploads/2021/04/WGIC-Report-2021-01-Geospatial-AI-ML-April-2021.pdf.

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. Environ. Res. Lett. 15 (2), 024019. https://doi.org/10.1088/1748-9326/ab68ac.

Xing, J., Sieber, R., 2021. September 1). Integrating XAI and GeoAI. GIScience 2021 Short Paper Proceedings.

Yan, X., Zang, Z., Jiang, Y., Shi, W., Guo, Y., Li, D., Zhao, C., Husi, L., 2021. A Spatial-Temporal Interpretable Deep Learning Model for improving interpretability and predictive accuracy of satellite-based PM. Environ. Pollut. 273 https://doi.org/10.1016/j.envpol.2021.116459.

Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks (pp. 818–833). https://doi.org/10.1007/978-3-319-10590-1_53.

Zhang, Y., Weng, Y., Lund, J., 2022. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. In: Diagnostics, Vol. 12, Issue 2. MDPI. https://doi.org/10.3390/diagnostics12020237.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., Youngblood, G.M., 2018. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. IEEE Conference on Computational Intelligence and Games (CIG) 2018, 1–8. https://doi.org/10.1109/CIG.2018.8490433.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE

Geosci. Remote Sens. Mag. 5 (4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307.