1
Towards Explainable Artificial Intelligence

Wojciech Samek[1][0000−0002−6283−3265] and Klaus-Robert Müller[2,3,4][0000−0002−3861−7685]

[1] Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
wojciech.samek@hhi.fraunhofer.de
[2] Technische Universität Berlin, 10587 Berlin, Germany
[3] Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea
[4] Max Planck Institute for Informatics, Saarbrücken 66123, Germany
klaus-robert.mueller@tu-berlin.de

Abstract. In recent years, machine learning (ML) has become a key enabling technology for the sciences and industry. Especially through improvements in methodology, the availability of large databases and increased computational power, today's ML algorithms are able to achieve excellent performance (at times even exceeding the human level) on an increasing number of complex tasks. Deep learning models are at the forefront of this development. However, due to their nested non-linear structure, these powerful models have been generally considered "black boxes", not providing any information about what exactly makes them arrive at their predictions. Since in many applications, e.g., in the medical domain, such lack of transparency may be not acceptable, the development of methods for visualizing, explaining and interpreting deep learning models has recently attracted increasing attention. This introductory paper presents recent developments and applications in this field and makes a plea for a wider use of *explainable* learning algorithms in practice.

Keywords: Explainable Artificial Intelligence · Model Transparency · Deep Learning · Neural Networks · Interpretability

1.1 Introduction

Today's artificial intelligence (AI) systems based on machine learning excel in many fields. They not only outperform humans in complex visual tasks [16, 53] or strategic games [56, 83, 61], but also became an indispensable part of our every day lives, e.g., as intelligent cell phone cameras which can recognize and track faces [71], as online services which can analyze and translate written texts [11] or as consumer devices which can understand speech and generate human-like answers [90]. Moreover, machine learning and artificial intelligence have become indispensable tools in the sciences for tasks such as prediction, simulation or exploration [78, 15, 89, 92]. These immense successes of AI systems

mainly became possible through improvements in deep learning methodology [48, 47], the availability of large databases [17, 34] and computational gains obtained with powerful GPU cards [52].

Despite the revolutionary character of this technology, challenges still exist which slow down or even hinder the prevailance of AI in some applications. Examplar challenges are (1) the large complexity and high energy demands of current deep learning models [29], which hinder their deployment in resource restricted environments and devices, (2) the lack of robustness to adversarial attacks [55], which pose a severe security risk in application such as autonomous driving[5], and (3) the lack of transparency and explainability [76, 32, 18], which reduces the trust in and the verifiability of the decisions made by an AI system.

This paper focuses on the last challenge. It presents recent developments in the field of *explainable artificial intelligence* and aims to foster awareness for the advantages–and at times–also for the necessity of transparent decision making in practice. The historic second Go match between Lee Sedol and AlphaGo [82] nicely demonstrates the power of today's AI technology, and hints at its enormous potential for generating new knowledge from data when being accessible for human interpretation. In this match AlphaGo played a move, which was classified as "not a human move" by a renowned Go expert, but which was the deciding move for AlphaGo to win the game. AlphaGo did not explain the move, but the later play unveiled the intention behind its decision. With explainable AI it may be possible to also identify such novel patterns and strategies in domains like health, drug development or material sciences, moreover, the explanations will ideally let us comprehend the reasoning of the system and understand why the system has decided e.g. to classify a patient in a specific manner or associate certain properties with a new drug or material. This opens up innumerable possibilities for future research and may lead to new scientific insights.

The remainder of the paper is organized as follows. Section 1.2 discusses the need for transparency and trust in AI. Section 1.3 comments on the different types of explanations and their respective information content and use in practice. Recent techniques of explainable AI are briefly summarized in Section 1.4, including methods which rely on simple surrogate functions, frame explanation as an optimization problem, access the model's gradient or make use of the model's internal structure. The question of how to objectively evaluate the quality of explanations is addressed in Section 1.5. The paper concludes in Section 1.6 with a discussion on general challenges in the field of explainable AI.

1.2 Need for Transparency and Trust in AI

Black box AI systems have spread to many of today's applications. For machine learning models used, e.g., in consumer electronics or online translation services,

[5] The authors of [24] showed that deep models can be easily fooled by physical-world attacks. For instance, by putting specific stickers on a stop sign one can achieve that the stop sign is not recognized by the system anymore.

transparency and explainability are not a key requirement as long as the overall performance of these systems is good enough. But even if these systems fail, e.g., the cell phone camera does not recognize a person or the translation service produces grammatically wrong sentences, the consequences are rather unspectacular. Thus, the requirements for transparency and trust are rather low for these types of AI systems. In safety critical applications the situation is very different. Here, the intransparency of ML techniques may be a limiting or even disqualifying factor. Especially if single wrong decisions can result in danger to life and health of humans (e.g., autonomous driving, medical domain) or significant monetary losses (e.g., algorithmic trading), relying on a data-driven system whose reasoning is incomprehensible may not be an option. This intransparency is one reason why the adoption of machine learning to domains such as health is more cautious than the usage of these models in the consumer, e-commerce or entertainment industry.

In the following we discuss why the ability to explain the decision making of an AI system helps to establish trust and is of utmost importance, not only in medical or safety critical applications. We refer the reader to [91] for a discussion of the challenges of transparency.

1.2.1 Explanations Help to Find "Clever Hans" Predictors

Clever Hans was a horse that could supposedly count and that was considered a scientific sensation in the years around 1900. As it turned out later, Hans did not master the math but in about 90 percent of the cases, he was able to derive the correct answer from the questioner's reaction. Analogous behaviours have been recently observed in state-of-the-art AI systems [46]. Also here the algorithms have learned to use some spurious correlates in the training and test data and similarly to Hans predict right for the 'wrong' reason.

For instance, the authors of [44, 46] showed that the winning method of the PASCAL VOC competition [23] was often not detecting the object of interest, but was utilizing correlations or context in the data to correctly classify an image. It recognized boats by the presence of water and trains by the presence of rails in the image, moreover, it recognized horses by the presence of a copyright watermark[6]. The occurrence of the copyright tags in horse images is a clear artifact in the dataset, which had gone unnoticed to the organizers and participants of the challenge for many years. It can be assumed that nobody has systematically checked the thousands images in the dataset for this kind of artifacts (but even if someone did, such artifacts may be easily overlooked). Many other examples of "Clever Hans" predictors have been described in the literature. For instance, [73] show that current deep neural networks are distinguishing the classes "Wolf" and "Husky" mainly by the presence of snow in the image. The authors of [46] demonstrate that deep models overfit to padding artifacts when classifying airplanes, whereas [63] show that a model which was

[6] The PASCAL VOC images have been automatically crawled from flickr and especially the horse images were very often copyrighted with a watermark.

trained to distinguish between 1000 categories, has not learned dumbbells as an independent concept, but associates a dumbbell with the arm which lifts it. Such "Clever Hans" predictors perform well on their respective test sets, but will certainly fail if deployed to the real-world, where sailing boats may lie on a boat trailer, both wolves and huskies can be found in non-snow regions and horses do not have a copyright sign on them. However, if the AI system is a black box, it is very difficult to unmask such predictors. Explainability helps to detect these types of biases in the model or the data, moreover, it helps to understand the weaknesses of the AI system (even if it is not a "Clever Hans" predictor). In the extreme case, explanations allow to detect the classifier's misbehaviour (e.g., the focus on the copyright tag) from a single test image[7]. Since understanding the weaknesses of a system is the first step towards improving it, explanations are likely to become integral part of the training and validation process of future AI models.

1.2.2 Explanations Foster Trust and Verifiability

The ability to verify decisions of an AI system is very important to foster trust, both in situations where the AI system has a supportive role (e.g., medical diagnosis) and in situations where it practically takes the decisions (e.g., autonomous driving). In the former case, explanations provide extra information, which, e.g., help the medical expert to gain a comprehensive picture of the patient in order to take the best therapy decision. Similarly to a radiologist, who writes a detailed report explaining his findings, a supportive AI system should in detail explain its decisions rather than only providing the diagnosis to the medical expert. In cases where the AI system itself is deciding, it is even more critical to be able to comprehend the reasoning of the system in order to verify that it is not behaving like Clever Hans, but solves the problem in a robust and safe manner. Such verifications are required to build the necessary trust in every new technology.

There is also a social dimension of explanations. Explaining the rationale behind one's decisions is an important part of human interactions [30]. Explanations help to build trust in a relationship between humans, and should therefore be also part of human-machine interactions [3]. Explanations are not only an inevitable part of human learning and education (e.g., teacher explains solution to student), but also foster the acceptance of difficult decisions and are important for informed consent (e.g., doctor explaining therapy to patient). Thus, even if not providing additional information for verifying the decision, e.g., because the patient may have no medical knowledge, receiving explanations usually make us feel better as it integrates us into the decision-making process. An AI system which interacts with humans should therefore be explainable.

[7] Traditional methods to evaluate classifier performance require large test datasets.

1.2.3 Explanations are a Prerequisite for New Insights

AI systems have the potential to discover patterns in data, which are not accessible to the human expert. In the case of the Go game, these patterns can be new playing strategies [82]. In the case of scientific data, they can be unknown associations between genes and diseases [51], chemical compounds and material properties [68] or brain activations and cognitive states [49]. In the sciences, identifying these patterns, i.e., explaining and interpreting what features the AI system uses for predicting, is often more important than the prediction itself, because it unveils information about the biological, chemical or neural mechanisms and may lead to new scientific insights.

This necessity to explain and interpret the results has led to a strong dominance of linear models in scientific communities in the past (e.g. [42, 67]). Linear models are intrinsically interpretable and thus easily allow to extract the learned patterns. Only recently, it became possible to apply more powerful models such as deep neural networks without sacrificing interpretability. These explainable non-linear models have already attracted attention in domains such as neuroscience [87, 89, 20], health [33, 14, 40], autonomous driving [31], drug design [70] and physics [78, 72] and it can be expected that they will play a pivotal role in future scientific research.

1.2.4 Explanations are Part of the Legislation

The infiltration of AI systems into our daily lives poses a new challenge for the legislation. Legal and ethical questions regarding the responsibility of AI systems and their level of autonomy have recently received increased attention [21, 27]. But also anti-discrimination and fairness aspects have been widely discussed in the context of AI [28, 19]. The EU's General Data Protection Regulation (GDPR) has even added the *right to explanation* to the policy in Articles 13, 14 and 22, highlighting the importance of human-understandable interpretations derived from machine decisions. For instance, if a person is being rejected for a loan by the AI system of a bank, in principle, he or she has the right to know why the system has decided in this way, e.g., in order to make sure that the decision is compatible with the anti-discrimination law or other regulations. Although it is not yet clear how these legal requirements will be implemented in practice, one can be sure that transparency aspects will gain in importance as AI decisions will more and more affect our daily lives.

1.3 Different Facets of an Explanation

Recently proposed explanation techniques provide valuable information about the learned representations and the decision-making of an AI system. These explanations may differ in their information content, their recipient and their purpose. In the following we describe the different types of explanations and comment on their usefulness in practice.

1.3.1 Recipient

Different recipients may require explanations with different level of detail and with different information content. For instance, for users of AI technology it may be sufficient to obtain coarse explanations, which are easy to interpret, whereas AI researchers and developers would certainly prefer explanations, which give them deeper insights into the functioning of the model.

In the case of image classification such simple explanations could coarsely highlight image regions, which are regarded most relevant for the model. Several preprocessing steps, e.g., smoothing, filtering or contrast normalization, could be applied to further improve the visualization quality. Although discarding some information, such coarse explanations could help the ordinary user to foster trust in AI technology. On the other hand AI researchers and developers, who aim to improve the model, may require all the available information, including negative evidence, about the AI's decision in the highest resolution (e.g., pixel-wise explanations), because only this complete information gives detailed insights into the (mal)functioning of the model.

One can easily identify further groups of recipients, which are interested in different types of explanations. For instance, when applying AI to the medical domain these groups could be patients, doctors and institutions. An AI system which analyzes patient data could provide simple explanations to the patients, e.g., indicating too high blood sugar, while providing more elaborate explanations to the medical personal, e.g., unusual relation between different blood parameters. Furthermore, institutions such as hospitals or the FDA might be less interested in understanding the AI's decisions for individual patients, but would rather prefer to obtain global or aggregated explanations, i.e., patterns which the AI system has learned after analyzing many patients.

1.3.2 Information Content

Different types of explanation provide insights into different aspects of the model, ranging from information about the learned representations to the identification of distinct prediction strategies and the assessment of overall model behaviour. Depending on the recipient of the explanations and his or her intent, it may be advantageous to focus on one particular type of explanation. In the following we briefly describe four different types of explanations.

1. **Explaining learned representations**: This type of explanation aims to foster the understanding of the learned representations, e.g., neurons of a deep neural network. Recent work [12, 38] investigates the role of single neurons or group of neurons in encoding certain concepts. Other methods [84, 93, 64, 65] aim to interpret what the model has learned by building prototypes that are representative of the abstract learned concept. These methods, e.g., explain what the model has learned about the category "car" by generating a prototypical image of a car. Building such a prototype can be formulated within the activation maximization framework and has been shown to be

an effective tool for studying the internal representation of a deep neural network.

2. **Explaining individual predictions**: Other types of explanations provide information about individual predictions, e.g., heatmaps visualizing which pixels have been most relevant for the model to arrive at its decision [60] or heatmaps highlighting the most sensitive parts of an input [84]. Such explanations help to verify the predictions and establish trust in the correct functioning on the system. Layer-wise Relevance Propagation (LRP) [9, 58] provides a general framework for explaining individual predictions, i.e., it is applicable to various ML models, including neural networks [9], LSTMs [7], Fisher Vector classifiers [44] and Support Vector Machines [35]. Section 1.4 gives an overview over recently proposed methods for computing individual explanations.
3. **Explaining model behaviour**: This type of explanations go beyond the analysis of individual predictions towards a more general understanding of model behaviour, e.g., identification of distinct prediction strategies. The spectral relevance analysis (SpRAy) approach of [46] computes such meta explanations by clustering individual heatmaps. Each cluster then represents a particular prediction strategy learned by the model. For instance, the authors of [46] identify four clusters when classifying "horse" images with the Fisher Vector classifier [77] trained on the PASCAL VOC 2007 dataset [22], namely (1) detect the horse and rider, 2) detect a copyright tag in portrait oriented images, 3) detect wooden hurdles and other contextual elements of horseback riding, and 4) detect a copyright tag in landscape oriented images. Such explanations are useful for obtaining a global overview over the learned strategies and detecting "Clever Hans" predictors [46].
4. **Explaining with representative examples**: Another class of methods interpret classifiers by identifying representative training examples [41, 37]. This type of explanations can be useful for obtaining a better understanding of the training dataset and how it influences the model. Furthermore, these representative examples can potentially help to identify biases in the data and make the model more robust to variations of the training dataset.

1.3.3 Role

Besides the recipient and information content it is also important to consider the purpose of an explanation. Here we can distinguish two aspects, namely (1) the intent of the explanation method (what specific question does the explanation answer) and (2) our intent (what do we want to use the explanation for).

Explanations are relative and it makes a huge difference whether their intent is to explain the prediction as is (even if it is incorrect), whether they aim to visualize what the model "thinks" about a specific class (e.g., the true class) or whether they explain the prediction relative to another alternative ("why is this image classified as car and not as truck"). Methods such as LRP allow to answer all these different questions, moreover, they also allow to adjust the amount of positive and negative evidence in the explanations, i.e., visualize what speaks

for (positive evidence) and against (negative evidence) the prediction. Such fine-grained explanations foster the understanding of the classifier and the problem at hand.

Furthermore, there may be different goals for using the explanations beyond visualization and verification of the prediction. For instance, explanations can be potentially used to improve the model, e.g., by regularization [74]. Also since explanations provide information about the (relevant parts of the) model, they can be potentially used for model compression and pruning. Many other uses (certification of the model, legal use) of explanations can be thought of, but the details remain future work.

1.4 Methods of Explainable AI

This section gives an overview over different approaches to explainable AI, starting with techniques which are model-agnostic and rely on a simple surrogate function to explain the predictions. Then, we discuss methods which compute explanations by testing the model's response to local perturbations (e.g., by utilizing gradient information or by optimization). Subsequently, we present very efficient propagation-based explanation techniques which leverage the model's internal structure. Finally, we consider methods which go beyond individual explanations towards a meta-explanation of model behaviour.

This section is not meant to be a complete survey of explanation methods, but it rather summarizes the most important developments in this field. Some approaches to explainable AI, e.g., methods which find influencial examples [37], are not discussed in this section.

1.4.1 Explaining with Surrogates

Simple classifiers such as linear models or shallow decision trees are intrinsically interpretable, so that explaining its predictions becomes a trivial task. Complex classifiers such as deep neural networks or recurrent models on the other hand contain several layers of non-linear transformations, which largely complicates the task of finding what exactly makes them arrive at their predictions.

One approach to explain the predictions of complex models is to locally approximate them with a simple surrogate function, which is interpretable. A popular technique falling into this category is Local Interpretable Model-agnostic Explanations (LIME) [73]. This method samples in the neighborhood of the input of interest, evaluates the neural network at these points, and tries to fit the surrogate function such that it approximates the function of interest. If the input domain of the surrogate function is human-interpretable, then LIME can even explain decisions of a model which uses non-interpretable features. Since LIME is model agnostic, it can be applied to any classifier, even without knowing its internals, e.g., architecture or weights of a neural network classifier. One major drawback of LIME is its high computational complexity, e.g., for state-of-the-art models such as GoogleNet it requires several minutes for computing the explanation of a single prediction [45].

Similar to LIME which builds a model for locally approximating the function of interest, the SmoothGrad method [85] samples the neighborhood of the input to approximate the gradient. Also SmoothGrad does not leverage the internals of the model, however, it needs access to the gradients. Thus, it can also be regarded as a gradient-based explanation method.

1.4.2 Explaining with Local Perturbations

Another class of methods construct explanations by analyzing the model's response to local changes. This includes methods which utilize the gradient information as well as perturbation- and optimization-based approaches.

Explanation methods relying on the gradient of the function of interest [2] have a long history in machine learning. One example is the so-called Sensitivity Analysis (SA) [62, 10, 84]. Although being widely used as explanation methods, SA technically explains the change in prediction instead of the prediction itself. Furthermore, SA has been shown to suffer from fundamental problems such as gradient shattering and explanation discontinuities, and is therefore considered suboptimal for explanation of today's AI models [60]. Variants of Sensitivity Analysis exist which tackle some of these problems by locally averaging the gradients [85] or integrating them along a specific path [88].

Perturbation-based explanation methods [94, 97, 25] explicitly test the model's response to more general local perturbations. While the occlusion method of [94] measures the importance of input dimensions by masking parts of the input, the Prediction Difference Analysis (PDA) approach of [97] uses conditional sampling within the pixel neighborhood of an analyzed feature to effectively remove information. Both methods are model-agnostic, i.e., can be applied to any classifier, but are computationally not very efficient, because the function of interest (e.g., neural network) needs to be evaluated for all perturbations.

The meaningful perturbation method of [25, 26] is another model-agnostic technique to explaining with local perturbations. It regards explanation as a meta prediction task and applies optimization to synthesize the maximally informative explanations. The idea to formulate explanation as an optimization problem is also used by other methods. For instance, the methods [84, 93, 64] aim to interpret what the model has learned by building prototypes that are representative of the learned concept. These prototypes are computed within the activation maximization framework by searching for an input pattern that produces a maximum desired model response. Conceptually, activation maximization [64] is similar to the meaningful perturbation approach of [25]. While the latter finds a minimum perturbation of the data that makes $f(x)$ low, activation maximization finds a minimum perturbation of the gray image that makes $f(x)$ high. The costs of optimization can make these methods computationally very demanding.

1.4.3 Propagation-Based Approaches (Leveraging Structure)

Propagation-based approaches to explanation are not oblivious to the model which they explain, but rather integrate the internal structure of the model into the explanation process.

Layer-wise Relevance Propagation (LRP) [9, 58] is a propagation-based explanation framework, which is applicable to general neural network structures, including deep neural networks [13], LSTMs [7, 5], and Fisher Vector classifiers [44]. LRP explains individual decisions of a model by propagating the prediction from the output to the input using local redistribution rules. The propagation process can be theoretically embedded in the deep Taylor decomposition framework [59]. More recently, LRP was extended to a wider set of machine learning models, e.g., in clustering [36] or anomaly detection [35], by first transforming the model into a neural network ('neuralization') and then applying LRP to explain its predictions. The leveraging of the model structure together with the use of appropriate (theoretically-motivated) propagation rules, enables LRP to deliver good explanations at very low computational cost (one forward and one backward pass). Furthermore, the generality of the LRP framework allows also to express other recently proposed explanation techniques, e.g., [81, 95]. Since LRP does not rely on gradients, it does not suffer from problems such as gradient shattering and explanation discontinuities [60].

Other popular explanation methods leveraging the model's internal structure are Deconvolution [94] and Guided Backprogagation [86]. In contrast to LRP, these methods do not explain the prediction in the sense "how much did the input feature contribute to the prediction", but rather identify patterns in input space, that relate to the analyzed network output.

Many other explanation methods have been proposed in the literature which fall into the "leveraging structure" category. Some of these methods use heuristics to guide the redistribution process [79], others incorporate an optimization step into the propagation process [39]. The iNNvestigate toolbox [1] provides an efficient implementation for many of these propagation-based explanation methods.

1.4.4 Meta-Explanations

Finally, individual explanations can be aggregated and analyzed to identify general patterns of classifier behavior. A recently proposed method, spectral relevance analysis (SpRAy) [46], computes such meta explanations by clustering individual heatmaps. This approach allows to investigate the predictions strategies of the classifier on the whole dataset in a (semi-)automated manner and to systematically find weak points in models or training datasets.

Another type of meta-explanation aims to better understand the learned representations and to provide interpretations in terms of human-friendly concepts. For instance, the network dissection approach of [12, 96] evaluates the semantics of hidden units, i.e., quantify what concepts these neurons encode. Other recent work [38] provides explanations in terms of user-defined concepts and tests to which degree these concepts are important for the prediction.

1.5 Evaluating Quality of Explanations

The objective assessment of the quality of explanations is an active field of research. Many efforts have been made to define quality measures for heatmaps which explain individual predictions of an AI model. This section gives an overview over the proposed approaches.

A popular measure for heatmap quality is based on perturbation analysis [9, 75, 6]. The assumption of this evaluation metric is that the perturbation of relevant (according to the heatmap) input variables should lead to a steeper decline of the prediction score than the perturbation of input dimensions which are of lesser importance. Thus, the average decline of the prediction score after several rounds of perturbation (starting from the most relevant input variables) defines an objective measure of heatmap quality. If the explanation identifies the truly relevant input variables, then the decline should be large. The authors of [75] recommend to use untargeted perturbations (e.g., uniform noise) to allow fair comparison of different explanation methods. Although being very popular, it is clear that perturbation analysis can not be the only criterion to evaluate explanation quality, because one could easily design explanations techniques which would directly optimize this criterion. Examples are occlusion methods which were used in [94, 50], however, they have been shown to be inferior (according to other quality criteria) to explanation techniques such as LRP [8].

Other studies use the 'pointing game" [95] to evaluate the quality of a heatmap. The goal of this game is to evaluate the discriminativeness of the explanations for localizing target objects, i.e., it is compared if the most relevant point of the heatmap lies on the object of designated category. Thus, these measures assume that the AI model will focus most attention on the object of interest when classifying it, therefore this should be reflected in the explanation. However, this assumption may not always be true, e.g., "Clever Hans" predictors [46] may rather focus on context than of the object itself, irrespectively of the explanation method used. Thus, their explanations would be evaluated as poor quality according to this measure although they truly visualize the model's prediction strategy.

Task specific evaluation schemes have also been proposed in the literature. For example, [69] use the subject-verb agreement task to evaluate explanations of a NLP model. Here the model predicts a verb's number and the explanations verify if the most relevant word is indeed the correct subject or a noun with the predicted number. Other approaches to evaluation rely on human judgment [73, 66]. Such evaluation schemes relatively quickly become impractical if evaluating a larger number of explanations.

A recent study [8] proposes to objectively evaluate explanation for sequential data using ground truth information in a toy task. The idea of this evaluation metric is to add or subtract two numbers within an input sequence and measure the correlation between the relevances assigned to the elements of the sequence and the two input numbers. If the model is able to accurately perform the addition and subtraction task, then it must focus on these two numbers

(other numbers in the sequence are random) and this must be reflected in the explanation.

An alternative and indirect way to evaluate the quality of explanations is to use them for solving other tasks. The authors of [6] build document-level representations from word-level explanations. The performance of these document-level representations (e.g., in a classification task) reflect the quality of the word-level explanations. Another work [4] uses explanation for reinforcement learning. Many other functionally-grounded evaluations [18] could be conceived such as using explanations for compressing or pruning the neural network or training student models in a teacher-student scenario.

Lastly, another promising approach to evaluate explanations is based on the fulfillment of a certain axioms [80, 88, 54, 60, 57]. Axioms are properties of an explanation that are considered to be necessary and should therefore be fulfilled. Proposed axioms include relevance conservation [60], explanation continuity [60], sensitivity [88] and implementation invariance [88]. In contrast to the other quality measures discussed in this section, the fulfillment or non-fulfillment of certain axioms can be often shown analytically, i.e., does not require empirical evaluations.

1.6 Challenges and Open Questions

Although significant progress has been made in the field of explainable AI in the last years, challenges still exist both on the methods and theory side as well as regarding the way explanations are used in practice. Researchers have already started working on some of these challenges, e.g., the objective evaluation of explanation quality or the use of explanations beyond visualization. Other open questions, especially those concerning the theory, are more fundamental and more time will be required to give satisfactory answers to them.

Explanation methods allow us to gain insights into the functioning of the AI model. Yet, these methods are still limited in several ways. First, heatmaps computed with today's explanation methods visualize "first-order" information, i.e., they show which input features have been identified as being relevant for the prediction. However, the relation between these features, e.g., whether they are important on their own or only whether they occur together, remains unclear. Understanding these relations is important in many applications, e.g., in the neurosciences such higher-order explanations could help us to identify groups of brain regions which act together when solving a specific task (brain networks) rather than just identifying important single voxels.

Another limitation is the low abstraction level of explanations. Heatmaps show that particular pixels are important without relating these relevance values to more abstract concepts such as the objects or the scene displayed in the image. Humans need to interpret the explanations to make sense them and to understand the model's behaviour. This interpretation step can be difficult and erroneous. Meta-explanations which aggregate evidence from these low-level heatmaps and explain the model's behaviour on a more abstract, more human

understandable level, are desirable. Recently, first approaches to aggregate low-level explanations [46] and quantify the semantics of neural representations [12] have been proposed. The construction of more advanced meta-explanations is a rewarding topic for future research.

Since the recipient of explanations is ultimately the human user, the use of explanations in human-machine interaction is an important future research topic. Some works (e.g., [43]) have already started to investigate human factors in explainable AI. Constructing explanations with the right user focus, i.e., asking the right questions in the right way, is a prerequisite to successful human-machine interaction. However, the optimization of explanations for optimal human usage is still a challenge which needs further study.

A theory of explainable AI, with a formal and universally agreed definition of what explanations are, is lacking. Some works made a first step towards this goal by developing mathematically well-founded explanation methods. For instance, the authors of [59] approach the explanation problem by integrating it into the theoretical framework of Taylor decomposition. The axiomatic approaches [88, 54, 60] constitute another promising direction towards the goal of developing a general theory of explainable AI.

Finally, the use of explanations beyond visualization is a wide open challenge. Future work will show how to integrate explanations into a larger optimization process in order to, e.g., improve the model's performance or reduce its complexity.

Acknowledgements. This work was supported by the German Ministry for Education and Research as Berlin Big Data Centre (01IS14013A), Berlin Center for Machine Learning (01IS18037I) and TraMeExCo (01IS18056A). Partial funding by DFG is acknowledged (EXC 2046/1, project-ID: 390685689). This work was also supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451, No. 2017-0-01779).

References

1. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.R., Dähne, S., Kindermans, P.J.: iNNvestigate neural networks!. Journal of Machine Learning Research **20**(93), 1–8 (2019)
2. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Gradient-based attribution methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
3. Antunes, P., Herskovic, V., Ochoa, S.F., Pino, J.A.: Structuring dimensions for collaborative systems evaluation. ACM Computing Surveys (CSUR) **44**(2), 8 (2012)
4. Arjona-Medina, J.A., Gillhofer, M., Widrich, M., Unterthiner, T., Hochreiter, S.: RUDDER: Return Decomposition for Delayed Rewards. arXiv preprint arXiv:1806.07857 (2018)
5. Arras, L., Arjona-Medina, J., Gillhofer, M., Widrich, M., Montavon, G., Müller, K.R., Hochreiter, S., Samek, W.: Explaining and interpreting LSTMs with LRP. In:

Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
6. Arras, L., Horn, F., Montavon, G., Müller, K.R., Samek, W.: "What is relevant in a text document?": An interpretable machine learning approach. PLoS ONE **12**(8), e0181142 (2017)
7. Arras, L., Montavon, G., Müller, K.R., Samek, W.: Explaining recurrent neural network predictions in sentiment analysis. In: EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA). pp. 159–168 (2017)
8. Arras, L., Osman, A., Müller, K.R., Samek, W.: Evaluating recurrent neural network explanations. In: ACL'19 Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2019)
9. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
10. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. Journal of Machine Learning Research **11**, 1803–1831 (2010)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (ICLR). (2015)
12. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6541–6549 (2017)
13. Binder, A., Bach, S., Montavon, G., Müller, K.R., Samek, W.: Layer-wise relevance propagation for deep neural network architectures. In: Information Science and Applications (ICISA), pp. 913–922 (2016)
14. Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., Stenzinger, A., Parlow, L., Budczies, J., Goeppert, B., et al.: Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. arXiv preprint arXiv:1805.11178 (2018)
15. Chmiela, S., Sauceda, H.E., Müller, K.R., Tkatchenko, A.: Towards exact molecular dynamics simulations with machine-learned force fields. Nature Communications **9**(1), 3887 (2018)
16. Cireşan, D., Meier, U., Masci, J., Schmidhuber, J.: A committee of neural networks for traffic sign classification. In: International Joint Conference on Neural Networks (IJCNN). pp. 1918–1921 (2011)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
18. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
19. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., Wood, A.: Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134 (2017)
20. Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A.U., Ruprecht, K., Giess, R.M., Kuchling, J., Asseyer, S., Weygandt, M., Haynes, J.D., et al.: Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation. arXiv preprint arXiv:1904.08771 (2019)

21. European Commission's High-Level Expert Group: Draft ethics guidelines for trustworthy AI. European Commission (2019)
22. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015)
23. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
24. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945 (2017)
25. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: IEEE International Conference on Computer Vision (CVPR). pp. 3429–3437 (2017)
26. Fong, R., Vedaldi, A.: Explanations for attributing deep neural network predictions. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
27. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**(3), 50–57 (2017)
28. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: From discrimination discovery to fairness-aware data mining. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2125–2126 (2016)
29. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: Advances in Neural Information Processing Systems (NIPS). pp. 1135–1143 (2015)
30. Heath, R.L., Bryant, J.: Human communication theory and research: Concepts, contexts, and challenges. Routledge (2013)
31. Hofmarcher, M., Unterthiner, T., Arjona-Medina, J., Klambauer, G., Hochreiter, S., Nessler, B.: Visual scene understanding for autonomous driving using semantic segmentation. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
32. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainabilty of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery p. e1312 (2019)
33. Horst, F., Lapuschkin, S., Samek, W., Müller, K.R., Schöllhorn, W.I.: Explaining the unique nature of individual gait patterns with deep learning. Scientific Reports **9**, 2391 (2019)
34. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 1725–1732 (2014)
35. Kauffmann, J., Müller, K.R., Montavon, G.: Towards explaining anomalies: A deep Taylor decomposition of one-class models. arXiv preprint arXiv:1805.06230 (2018)
36. Kauffmann, J., Esders, M., Montavon, G., Samek, W., Müller, K.R.: From clustering to cluster explanations via neural networks. arXiv preprint arXiv:1906.07633 (2019)
37. Khanna, R., Kim, B., Ghosh, J., Koyejo, O.: Interpreting black box predictions using fisher kernels. arXiv preprint arXiv:1810.10118 (2018)
38. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning (ICML). pp. 2673–2682, (2018)

39. Kindermans, P.J., Schütt, K.T., Alber, M., Müller, K.R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. In: International Conference on Learning Representations (ICLR). (2018)
40. Klauschen, F., Müller, K.R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., de Maria, S., Badve, S., et al.: Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Seminars in Cancer Biology **52**(2), 151–157 (2018)
41. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International Conference on Machine Learning (ICML). pp. 1885–1894 (2017)
42. Kriegeskorte, N., Goebel, R., Bandettini, P.: Information-based functional brain mapping. Proceedings of the National Academy of Sciences **103**(10), 3863–3868 (2006)
43. Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., Doshi-Velez, F.: An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006 (2019)
44. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W.: Analyzing classifiers: Fisher vectors and deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2912–2920 (2016)
45. Lapuschkin, S.: Opening the Machine Learning Black Box with Layer-wise Relevance Propagation. Ph.D. thesis, Technische Universität Berlin (2019)
46. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature Communications **10**, 1096 (2019)
47. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
48. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R.: Efficient backprop. In: Neural networks: Tricks of the trade, pp. 9–48. Springer (2012)
49. Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.R.: Introduction to machine learning for brain imaging. Neuroimage **56**(2), 387–399 (2011)
50. Li, J., Monroe, W., Jurafsky, D.: Understanding Neural Networks through Representation Erasure. arXiv preprint arXiv:1612.08220 (2016)
51. Libbrecht, M.W., Noble, W.S.: Machine learning applications in genetics and genomics. Nature Reviews Genetics **16**(6), 321 (2015)
52. Lindholm, E., Nickolls, J., Oberman, S., Montrym, J.: Nvidia tesla: A unified graphics and computing architecture. IEEE Micro **28**(2), 39–55 (2008)
53. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with GaussianFace. In: 29th AAAI Conference on Artificial Intelligence. pp. 3811–3819 (2015)
54. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NIPS). pp. 4765–4774 (2017)
55. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR). (2018)
56. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
57. Montavon, G.: Gradient-based vs. propagation-based explanations: An axiomatic comparison. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)

58. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: An overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
59. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
60. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
61. Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., et al.: Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. Science **356**(6337), 508–513 (2017)
62. Morch, N., Kjems, U., Hansen, L.K., Svarer, C., Law, I., Lautrup, B., Strother, S., Rehm, K.: Visualization of neural networks using saliency maps. In: International Conference on Neural Networks (ICNN). vol. 4, pp. 2085–2090 (1995)
63. Mordvintsev, A., Olah, C., Tyka, M.: Inceptionism: Going deeper into neural networks (2015)
64. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 3387–3395 (2016)
65. Nguyen, A., Yosinski, J., Clune, J.: Understanding neural networks via feature visualization: A survey. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
66. Nguyen, D.: Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 1069–1078 (2018)
67. Phinyomark, A., Petri, G., Ibáñez-Marcelo, E., Osis, S.T., Ferber, R.: Analysis of big data in gait biomechanics: Current trends and future directions. Journal of Medical and Biological Engineering **38**(2), 244–260 (2018)
68. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., Ramprasad, R.: Accelerating materials property predictions using machine learning. Scientific Reports **3**, 2810 (2013)
69. Poerner, N., Roth, B., Schütze, H.: Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In: 56th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 340–350 (2018)
70. Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S., Unterthiner, T.: Interpretable deep learning in drug discovery. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
71. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
72. Reyes, E., Estévez, P.A., Reyes, I., Cabrera-Vives, G., Huijse, P., Carrasco, R., Forster, F.: Enhanced rotational invariant convolutional neural network for supernovae detection. In: International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2018)
73. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)

74. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: 26th International Joint Conferences on Artificial Intelligence (IJCAI). pp. 2662–2670 (2017)
75. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems **28**(11), 2660–2673 (2017)
76. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services **1**(1), 39–48 (2018)
77. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.J.: Image classification with the Fisher vector: Theory and practice. International Journal of Computer Vision **105**(3), 222–245 (2013)
78. Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A.: Quantum-chemical insights from deep tensor neural networks. Nature Communications **8**, 13890 (2017)
79. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (CVPR). pp. 618–626 (2017)
80. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
81. Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences. arXiv preprint arXiv:1704.02685 (2017)
82. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., et al.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
83. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of Go without human knowledge. Nature **550**(7676), 354–359 (2017)
84. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR Workshop. (2014)
85. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
86. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR Workshop. (2015)
87. Sturm, I., Lapuschkin, S., Samek, W., Müller, K.R.: Interpretable deep neural networks for single-trial eeg classification. Journal of Neuroscience Methods **274**, 141–145 (2016)
88. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (ICML). pp. 3319–3328 (2017)
89. Thomas, A.W., Heekeren, H.R., Müller, K.R., Samek, W.: Analyzing neuroimaging data through recurrent deep learning models. arXiv preprint arXiv:1810.09945 (2018)
90. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. SSW **125** (2016)
91. Weller, A.: Transparency: Motivations and Challenges. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)

92. Wu, D., Wang, L., Zhang, P.: Solving statistical mechanics using variational autoregressive networks. Physical Review Letters **122**(8), 080602 (2019)
93. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
94. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference Computer Vision (ECCV). pp. 818–833 (2014)
95. Zhang, J., Lin, Z.L., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: European Conference on Computer Vision (ECCV). pp. 543–559 (2016)
96. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Comparing the interpretability of deep networks via network dissection. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science **11700**, Springer (2019)
97. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. In: International Conference on Learning Representations (ICLR). (2017)