A Survey on Explainability: Why should we believe the accuracy of a model?

PRARTHANA DUTTA, NARESH BABU MUPPALANENI, and RIPON PATGIRI*, National Institute of Technology silchar, India

The world has been evolving with new technologies and advances day-by-day. With the advent of various learning technologies in every field, the research community is able to provide solution in every aspect of life with the applications of Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision, etc. However, with such high achievements, it is found to lag behind the ability to provide explanation against its prediction. The current situation is such that these modern technologies are able to predict and decide upon various cases more accurately and speedily than a human, but failed to provide an answer when the question of why to trust its prediction is put forward. In order to attain a deeper understanding into this rising trend, we explore a very recent and talked-about novel contribution which provides rich insight on a prediction being made – "Explainability." The main premise of this survey is to provide an overview for researches explored in the domain and obtain an idea of the current scenario along with the advancements published to-date in this field. This survey is intended to provide a comprehensive background of the broad spectrum of Explainability.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

Additional Key Words and Phrases: Artificial Intelligence, Explainability, Deep Learning, Machine Learning.

1 INTRODUCTION

Advancements in technology in every field field of life have been witnessed by researchers across generations. John McCarthy in 1956 coined the term "Artificial Intelligence" (AI), and since then, it has been evolving as an elusive subject of concern in many research activities. AI has been witnessing their utility in various fields since decades and have been achieving desired and satisfactory achievements. There have been tremendous improvements in a wide range of domains such as commercial platforms, medical, marketing, etc. This has paved the way such that no human intervention is needed in most aspects of decision taking. With more and more advancements in the learning technologies and the models, the decisions or predictions made by these systems are accurate approximately 100%. Also, machine learning algorithms can correctly predict and its accuracy may rise to 100% in some favorable cases. Patgiri *et al.* [38] reports accuracy of 100% in conventional machine learning algorithms. Even though with such high and satisfactory accuracies,

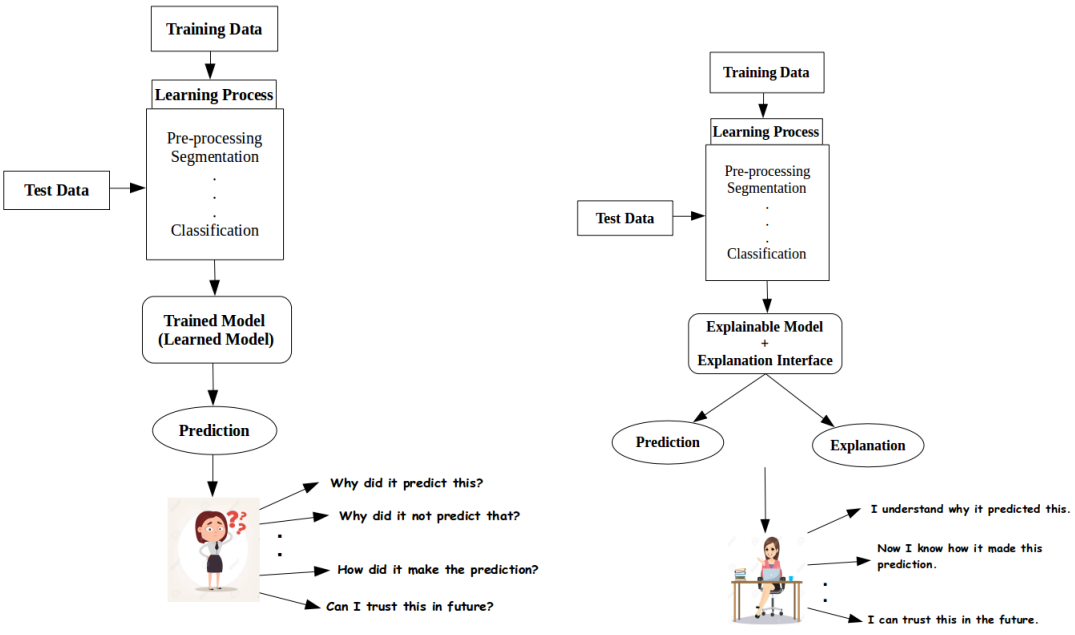
*Corresponding author

Authors' address: Prarthana Dutta, prarthana.dutta01@gmail.com; Naresh Babu Muppalaneni, nareshmuppalaneni@gmail.com; Ripon Patgiri, ripon@cse.nits.ac.in, Department of Computer Science & Engineering, National Institute of Technology silchar, Silchar, Assam, India, 788010.

these technologies may face trust issues related to the model and the predictions [38, 40]. Hence, the main risk that is associated with these predictions is that even though the model was able to predict accurately for all the test cases, it might not work in the same accurate manner when the model is left in the wild for their respective domains to track and make life changing decisions (some of them are discussed in other sections). Thus a gap is found to exist between the prediction or decision made by an intelligent system and the reason associated with these decisions or predictions. The end users in most cases remain in a confused state of mind having various curiosities about the model and its prediction. A pictorial view of such a scenario is shown in Figure 1a. For this, the researchers have been thinking to take their research one step ahead by thinking if the system gave its end user a valid reason for making such decisions along with the predictions. They developed this idea that along with making the predictions, if the intelligent systems (making these predictions) are also able to give a proper explanation of its prediction, it would definitely prove to be much easier and more meaningful for end users to rely and trust the systems and take further actions based on this explanation. Thus, bridging the gap between the prediction and the reason for such prediction for humans to understand and interpret better is what we call an explainable system and the mechanism inbuilt in the model for explaining its decision is called as **Explainability**. Despite handling the trust issues associated with the model and its prediction, explainability also found its importance in other aspects such as providing understandability and transparency into the working algorithm and the model employed. The usability and application of explainability can be seen in a wide range of domains such as Healthcare, Education, Business, Military, etc., discussed in Section 2.1.

Since the discovery of various prediction models and algorithms, one never gains insights into the internal structure and the working of the model which led to the arrival of the desired prediction. It is mostly imagined as a *black-box* with no understanding about its working. With the ability of peeping into this black-box and exploring its contents and working, one can explain the reason behind any prediction made by the model. In later times one can also try to modify its content, if necessary, for smooth running, prediction making and explaining (Section 4). Figure 1a describes a general architecture of a learning technology in making a prediction. The training data is fed into the learning process or architecture and the output induces a model of the training data. This model is then used to make predictions for new test data. The problem with this prediction is that, it is not accompanied with any justification or explanation for the specific prediction it made. The end user at times may be left in a confused state of mind dealing with various questions about the model and its prediction. They are unaware and unable to view anything that has been learnt within the model. The explainability feature comes into play in these scenarios by trying to generate a new learning process whose output is justified with the prediction. It allows the user to interrogate the internal model working and obtain justification against the prediction. This can be visualized from Figure 1b where the end user is able to gain insights into the working of the model via the explanation provided by the explainable model and the interface. The user now becomes aware of what the model has learnt.

The motivation behind this survey is to provide new interested researchers a clear picture of the current scenario of explainability in learning technology. This survey is intended to bring into limelight the recent trends which the research community is following to deal with explainability. This overview is organized as: Section 2 briefs on Explainability in Machine Learning where we explored about how the term 'explainability' is defined by various researchers and also how the *black-box problem* was explored. This section also brings about the requirement of explainability across various domains and succinctly discusses the trade-off between explainability and accuracy. Section 3 discusses the various techniques and frameworks employed in explainability. Section 4 surveys about the importance of the cognitive science in designing an explainable model.



(a) Conventional learning (b) Learning methods with Explainability

Fig. 1. (a) A general learning technology which provides the user with state-of-the-art performance measures but is unable to fulfil the users queries and leaves the user in a confused state of mind. (b) An explainable technology which is unified with an explainable model and interface along with the prediction

2 EXPLAINABILITY IN MACHINE LEARNING

The term eXplainable Artificial Intelligence is usually abbreviated as **XAI**. The term was first formulated by Lent and Mancuso in 2004 [53]. Before that it was addressed as "black-box". The main aim behind the development of such a technique is twofold — Firstly, it provides transparency against the working models and algorithms in a way to gain insights into the black-box. Secondly, it aids in converting an untrustworthy model into a trustworthy one.

Digging deep into the matter, it is found that Explainability is not a recent topic that the researchers have been focusing upon. It has been discussed for many years and have recently gained more attention and importance due to their ability to explain the predictions made by the intelligent systems which is indeed a major concern for the study. Various researchers define Explainability in a number of ways — Ribeiro *et al.* [40] in 2016 defined explainability as "presenting textual or visual artifacts that provide a qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction." According to Guidotti *et al.* [12] "an explanation is an Interface between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans." Gilpin *et al.* defined it in terms of " models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions." [10]. According to Tim Miller [33], Explainable Artificial Intelligence (XAI) refers to "an explanatory agent revealing underlying causes to its or another agent's decision making." Recently

Arrieta *et al.* [4] have also defined it as "Given an audience an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear to understand."

The survey of Adadi *et al.* [1] highlights the main concepts and terms related to XAI and provides us with a structured view of these concepts. The survey also highlights the need for an explainable model so that it can be utilized in every field possible.

In a very recent and elaborate survey published in 2019, by Arrieta *et al.* [4], the authors discuss the rising trend of explainability in Machine Learning. According to the authors, explainability could be achieved successfully either through models that are transparent, i.e. they are understandable by themselves, or by application of Post-hoc techniques to understand them. The self-understandable models include Logistic/Linear Regression, Decision Trees, K-Nearest Neighbors, etc. While coming to the Post-hoc technique, the authors made an elaborate and hierarchical categorization constituting of the Model-Agnostic and Model-Specific techniques with their respective taxonomies. We narrowed down our survey by analyzing upon the Model-Specific Learning Technologies for Deep Learning Models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), etc. Studies show that Deep learning methods such as CNN, RNN etc. have been (proven their utility) achieving great heights in predicting the accuracies of complex data. Questioning on how these models generated the outputs or predictions remain unanswered. Without understanding how they arrived at such solutions, it won't be relevant to employ the model out in the wild.

2.1 Need for Explainability in various domains

Though the AI is able to help the society in real life decision making problems, there are situations witnessed where the AI is seen to fail and is disowned in many situations. This calls for an urgent need for a valid and trustworthy explainable system that could be trusted blindly by the users and organizations. Some of the scenarios where one needs an explainable model while taking life changing decisions and have faced failure due to lack of explainability are shown in Table 1.

Table 1. Explainability studies carried out in Deep Learning models for providing visual and feature relevance explanation

Domain	Situation	Observation	Cause of Failure
Education [28]	Selecting candidates for interview	Practicing sexual and racial discrimination.	Inferred information from the last name and date of birth of the candidates.
Military [9, 17]	Separating friendly and enemy tanks	Poor accuracy on the test data	The model was trained with friendly photos were taken on sunny days while enemy photos were taking on overcast days.
Healthcare [7]	Pneumonia risk prediction	Pneumonia risk patients with asthmatic history was nopt considered as a serious condition; which is clinically in correct.	Neural Network inferred that patients with asthma has low risk of dying and can be treated as outpatients.
Husky Vs Wolf [40]	Distinguishing between images of wolves and Eskimo Dogs	Wolf not in a snowy background is predicted as Husky and Husky in a snowy background is predicted as Wolf.	Prediction was solely based on the presence or absence of snow in the background.
Transportation [32]	Self-driving Car	Self-driving Uber killed a women in Arizona	Misclassified a women (object) as a plastic bag or tumbler in the air .

In the 1982s, St George's Hospital Medical School used a computer application for screening of applicants for interview. But the computer program is found to have discriminated against women candidates and non-European applicants and had less chance of selection for the interview [28].

This discrimination came into notice later on when they learned that the system inferred that women are more likely to ask for time off work due to their family bonds. Again the program also deduced that non-Europeans may not have sufficient command over the English language to practice medicine. But the selection of candidates on the basis of these factors are not valid and hence the system was abandoned by the authority. The need for an explainable model is being witnessed in an incident related to the military domain [9, 17]. Alex A. Freitas tried to evaluate the performance of classification models based on predictive accuracy and comprehensibility in 2014. He performed a simple experiment of classifying images of takes as enemy tanks and friendly tanks. The predictive accuracy performance of the ANN was observed to be high and so they decided to set the model out for practical implementation. But it did not perform well in classifying the enemy and friendly tanks. Later on it was reported that all the friendly photos on the training data were taken on sunny days while all the enemy photos were taken on overcast days. Hence the learning model inferred its classification based on the color of the sky on the photos. An explainable model is obvious to explain that it discriminated the photos based on the color of the sky and not on the relevant features of the images. With this explanation against the classification, the authority could have decided whether or not to employ the model for practical applications.

A similar example was stated by Ribeiro *et al.* while trying to figure out the importance of explanation in matters of trusting the model [39]. They trained a classifier to distinguish between images of Wolves and Huskins. At first they used a logistic regression classifier to classify images of wolves and huskies. The images were picked up intentionally such that all images of wolves are taken on a snowy background while husky images are taken without snow in the background. This *bad* classifier was set to be tested on 10 images under two scenarios — First without explanation and second with explanation. These test images also intentionally contained one wolf and husky image without and with snowy background respectively. So the classifier classified 8 of the test samples correctly while the two images of the wolf and husky are misclassified i.e. wolf without the snowy background was classified as husky and husky with the snowy background was classified as a wolf. In the second phase of the experiment, explanations are provided against the prediction. Evaluators are asked the question whether the model could be trusted or not. The result is such that, without the explanation provided, 10 among 27 trusted the model as making the correct prediction. On the other hand, after the explanation were provided, only 3 among 27 trusted the bad model. So the importance of explanation in a prediction could be visualized from this experiment.

To investigate the utility of Machine Learning in medical applications, Caruana *et al.* tried to conduct a study on pneumonia risk prediction in order to predict the probability of death in pneumonia risk patients [7]. It was to determine whether patients with pneumonia risk are required to be treated as serious and admitted to the hospital or treated as outpatients. It turned out that while training the model with the medical history of the patients, it incorrectly learned the rule which inferred: $Asthma(patient) \implies Low - Risk(patient)$, i.e. patients with asthma in their medical history and having pneumonia have a low-risk of dying due to pneumonia. According to this rule, those patients with pneumonia and having a medical history of asthma are in a low-risk of death and need not be admitted to the hospital. But, clinically this pose a serious threat as pneumonia patients with a medical history of asthma are in a serious condition and required to be admitted to the hospital immediately and needed special medical attention. As a result, the medical authorities decided to abandon the model. Self-driving vehicles are being welcomed in Arizona since December 2006. But it proved to be a failure in the year 2018 while taking a test drive in automated mode on the streets of Tempe, Arizona [32]. A Volvo SUV is reported to have hit and killed a 49 year old women walking on her bicycle across the street in Tempe. After this incident the testing of self-driving vehicles have been suspended in the US and Canada. The car's autonomous mode failed to sense and detect the pedestrian, even though the vehicle was trained and designed to detect

pedestrians, cyclists, etc. even in darkness. This was reported as failure in classification by experts. It was found that the car's software misclassified the pedestrian to be a plastic bag or a tumbler and treated it as such.

2.2 Explainability and Accuracy

Debugging the intelligent system is an integral part of research. This help developers in a number of ways — correct the mistakes and shortcomings in the system, trust the system, etc. In [21] authors Kulesza *et al.*, performed an illusive study on how the models equipped with explainable models should explain their prediction the users. They analysed their study with two important aspects of explanation — 'Soundness' and 'Completeness'. They interpreted Soundness as the truthfulness associated with each component of the system and Completeness as the ability to describe all the intrinsic systems. In their attempt to answer some of the research questions, they performed the experiment on a music recommendation system under four different circumstances, described as High Soundness and High Completeness (HH), Medium Soundness and Medium Completeness (MM), High Soundness and Low Completeness (HSLC) and Low Soundness and High Completeness (LSHC). They tried to answer queries relating to impact of soundness and completeness on the model's mental health; beneficial information; obstacles; cost-benefit trade off and trust. Most complete models are proven to be beneficial for the mental health of the model. Complete systems also proven to be associated with low cost and high benefit incurred. Complete systems are also associated with the trust associated with the model. Hence, the end result indicated that completeness is inferior to soundness when the model tries to explain how it arrived at the prediction. With the development of the concept of explainability as a new learning technology, various myths and misconceptions are also associated with the concept. Most of these remain unobserved, which may pose a threat to the conclusions made by the model. Most of the misconception, as highlighted by many research studies, including the Defense Advanced Research Project Agency (DARPA) [13, 52] and others, such as [41], is the trade-off existing between the accuracy computed by the model along with the explanation provided by it. The accuracy of a model is, in most cases, independent of the complexity of the model. Though most of the research community believes that more complex structured algorithms produce more accurate results, i.e. the more complex is the model, the higher is the accuracy. This is not always true when one has to deal with the structured and meaningful data. When dealing with such data, both the simple and complex structured models show similar accuracy; but their explainability may vary.

Coming to the Computer Vision applications, the Deep Learning models show higher performance with less ability to provide explainability. But there are also situations witnessed where explainability and accuracy in performance are found to be directly proportional in different Machine Learning algorithms with slight variation in their performance [14]. To visualize these scenarios where the DARPA project for XAI has provided a graphical view of the relationship of the Machine Learning Models towards explainability and their individual performance. From Figure 2 it is clear that though the Machine Learning models such as Deep Learning, Random Forests, Support Vector Machines has high accuracy and less explainability; while the other models such as Bayesian Belief Networks, Decision Trees, etc. can provide higher explainability but their accuracy does not seem to achieve heights as compared with the other ML models. So with XAI, the researchers need to develop or modify their models such that they attain higher accuracy along with providing a good explanation measure. Thus, one needs to increase explainability such that the accuracy results are not sacrificed.

Again, it may not always the case that the explanations provided by the models can be blindly trusted. It may sometimes go wrong when the question of providing correct explanation is concerned.

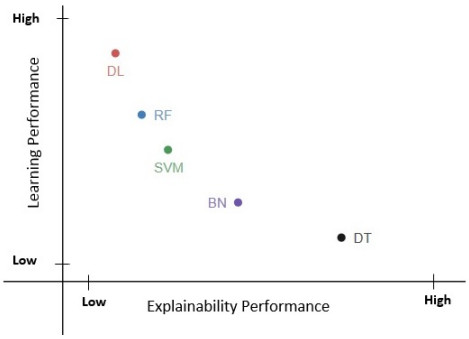


Fig. 2. Relationship between Explainability and Performance Measure among the various Learning Models. This is an approximation graph presented by the DARPA project on their study on eXplainable Artificial Intelligence. Source [13].

So the models designed may be expected to provide explanation a majority portion of the original mode.

3 MATERIALS AND METHODS

There are a wide variety of materials and methods provided by the research community in the context of XAI. These materials and methods are dependent on the application domains of the models being used. There are various explanation techniques and frameworks which are developed by researchers across the globe to help them attain desired results. We have decided to put special emphasis on only a few application domain of Machine Learning that focuses mainly on visual explanation and feature relevance explanations. These explanation techniques and frameworks are explored in Section 3.1 and Section 3.2 respectively. The tables provided in these sections provide a brief idea of the correct approaches and frameworks employed in the relevant fields of explainability providing insights into the approaches, visualization techniques, datasets, etc.

3.1 Explanation Techniques in XAI in Deep Learning

Many scholars have proposed a number of explanation methods and techniques for deep learning. One such taxonomy is provided recently by Arrieta *et al.*[4] which include — Explanation by Simplification, Feature Relevance Explanation, Visual Explanation, Local Explanation, Explanation of Example, Text Explanation, Visual Explanation, Architectural Modification, etc. Amongst them, we have chosen to survey on the Visual Explanation methods and Feature Relevance Explanation methods. Table 1 provides some of the approaches and visualization techniques employed in a few research works for providing explainability.

The alarming trends of achievements in classification techniques using Convolutional Neural Networks is known to all. With a view to know why and how the CNNs are able to provide such good accuracies, in 2013, Zeiler *et al.* studied on a visualization technique to peep deep into the internal layers of the CNN. They used a multi-layer Deconvolutional Network precisely thee 'deconv' model [57] to perform reverse mapping from the feature activations to the input space of pixel values to determine the specific input pixel value that resulted in the activation of the feature map [56]. So a deconv net is attached after every convnet layer. Simonyan *et al.*[47] devises two visualization techniques — Class Model visualization and Image-Specific Class Saliency visualization aimed at

Table 2. Explainability studies carried out in Deep Learning models for providing visual and feature relevance explanation

Author	Approach	Visualization techniques	Dataset
Zeiler *et al.* [56]	Fully-supervised convnet model	deconvnet	ImageNet 2012
Simonyan *et al.* [47]	deep ConvNet	Class Model and Image-Specific Class Saliency visualization	ILSVRC-2013
Bach *et al.* [5]	Taylor-type decomposition and Layer-wise relevance propagation	Heatmap	PASCAL VOC 2007, MNIST, ImageNet, synthetic image of geometric shapes.
Mahendran *et al.* [31]	Inverting Representation	Feed-forward & discriminatively trained CNN	ImageNet ILSVRC 2012
Li *et al.* [24]	Soft-spectral clustering and Bipartite matching		ILSVRC 2012
Goyal *et al.* [11]	Visual Question Answering Model	Guided back propagation and occlusion	VQA dataset
Nguyen *et al.* [37]	Activation Maximization	DGN-AM	ImageNet, MIT Places dataset
Selvaraju *et al.* [44]	Grad-CAM	Heatmap	PASCAL VOC 2007
Liu *et al.* [27]	CNNVis	Hybrid visualization	MNSIT, CIFAR – 10
Samek *et al.* [43]	Sensitivity Analysis(SA) and Layer-wise Relevance Propagation (LRP)	Heatmap	ILSVRC2012, 20Newsgroup, HMDB51

visualizing the class models and highlighting the areas of interest of the image of the particular class.

Aiming to view and understand the path followed by classifiers in arriving at the decisions, the research community contributed their ideas by exploring them in a number of ways. Bach, Binder *et al.* [5] proposed a general technique applicable to both neural network models and bag-of-words models for understanding the prediction by visualizing the pixels or prominent attributes that resulted in the prediction. According to the authors, while following the process of mapping the pixel values of the raw-image to the classifiers, there is no linear mapping among the various constituents of the process which leads to the absence of the explainability of the prediction being made. They proposed two pixels-wise decomposition approaches — Taylor-type decomposition and Layer-wise relevance propagation approach for the models, multi-layer neural networks and Bag of Words (BoW). Mahendran *et al.* developed a technique that provides explanation based on the visual content of the image [31]. They discover the method with a view to addressing the question of recovering an image based on a given encoding. The developed the *inverting representation* approach that is used to compute the inverse representation of the image. The experiment was conducted on the ImageNet [42] dataset and is found to work better as compared to recent trends [55]. While posing the question of whether different deep neural networks trained with different initialization, are able to perform *convergent learning*, Li, Yosinski *et al.* [24] is able to gain insight into deep learning models. They used a matching approach or a soft-spectral clustering approach in order to align the units from differed networks.

Goyal *et al.* [11] proposes two visualization techniques for the Visual Question Answering (VQA) (Lu *et al.*2015) [29] which provides an explanation to their prediction. The techniques are Guided Backpropagation [49] and Occlusion. These techniques are used to interpret and find out the pixels or words in an image or sentence which is mostly focused upon while answering the questions in the VQA model. The guided backpropagation and occlusion techniques are employed respectively to analyze important words or pixels (in a sentence or image respectively) and to observe changes

in the prediction probability by occluding the input respectively. Nguyen *et al.* [37] used Activation Maximization (AM) as a feature relevance explanation method and visualized it using a Deep Generator Network (DGN). The AM is used to synthesize the preferred input neuron. Authors Samek *et al.* [43] developed two approaches for visualizing the deep learning models. They are — **Sensitivity Analysis (SA)** [6], [48] for explaining a model prediction via the locally evaluated gradient of the model and **Layer-wise Relevance Propagation (LPR)** [5] for explaining the decision of the classifier using decomposition technique. They used three different classification approaches to obtain an explanation for the decisions being made — annotated images (ILSVRC2012 [42]), text document classification (20Newsgroup [36]) and activity recognition in videos (HMDB51 [18]).

3.2 Framework for Explainability in Deep Learning

Trying to understand why a model chosen by us for a particular computation, made such a prediction is an important topic of interest which have been evolving in recent times. Understanding the underlying reason behind such prediction may be able to incur something which is called trust. This is indeed an important aspect in the learning technologies using which one can take the decision whether or not to employ the model under consideration. Explaining a prediction refers to providing qualitative understanding of the relationship between the components of an instance and the prediction made by a model in the form of textual or visual artifacts. Developing a technique that can clearly explain the predictions predicted by a classifier is a challenging task. It addresses two of the most common problems i.e. trusting a model and trusting a prediction. Helps in picking up the most relevant classifier from among a set of classifiers. Accuracy measure may not be always suffice to explain a model prediction, one also needs explainability for such prediction. The main aim is to provide insight as to why the model gave us such prediction (trust related task). Also measures faithfulness of explanations. The various characteristics that are required to be fulfilled by a model to provide a proper explanation of its prediction require a well-designed framework that encompasses all the required properties explained in [4].

The authors developed an explainable model called LIME abbreviated as Local Interpretable Model-Agnostic Explanations, that can explain or detect features that contributed for such predictions. It thus provides a way of how humans can understand the working of the model and have faith in its prediction. It is inbuilt with characteristics such as interpretable, local fidelity, providing global perspective. In order to attain a deeper understanding into the model, the authors mainly focused on the aspect of **trust**. They developed this framework with a view to answering questions such as whether the predictions should be trusted by any user; whether the explanations provided by the framework could be used in general for selection of a model from a pair; etc. Towards answering the query about trusting a prediction and trusting the model, the authors presented the LIME and SP-LIME frameworks respectively. LIME was able to provide more than 90% recall on the books and DVDs datasets for the two interpretable classifiers, Decision Trees and Logistic regression classifier. The results indicate that the explanations provided by LIME are faithful to the model. At the same time it also proved to be a powerful tool for associating the trust in the predictions and are good in generalizing.

The SHAP model is SHarpely Additive exPlanations. They are used extensively to describe the Deep Learning models in particular. It is a model-agnostic unified framework which is used in interpreting predictions by unifying six methods altogether (LIME[39], DeepLIFT [46], layer-wise relevance propagation [5], shaply regression [26], shaply sampling [50], quantitative input influence feature attributions [8]). Sound LIME (SLIME) [35] is a version of the LIME framework that extends LIME to MCA systems. SLIME has the ability to produce three versions of explanations, including temporal, frequency and time-frequency segmentation. The framework is experimented on three

Table 3. Explainability studies carried out in Deep Learning models for providing visual and feature relevance explanation

Name	Year	Author	Description
LIME	2016	Riberio *et al.* [40]	Provide faithful explanation against the prediction made by a classifier. Addresses the "trusting a prediction" problem.
SP-LIME	2016	Riberio *et al.* [40]	Addresses the "trusting a model" problem.
SHAP	2017	Lundberg *et al.* [30]	Assigns value to each feature based on their importance while combining six existing methods into one new class.
SLIME	2017	Mishra *et al.* [35]	Extension of LIME to Music Component Analysis (MCA).
aLIME	2016	Ribeiro, Singh *et al.* [39]	Provide explanations using 'if-then' rule (rule-based explanations).
ASTRIDE	2017	Henelius *et al.* [15]	Aids in identifying interacting attributes for better explainability.
DeepLIFT	2016	Shrikumar *et al.* [45, 46]	Assigns a type of ranking to the determine the contribution score of the neuron.

singing voice systems — Decision Tree, Random Forest and Convolution neural Network and provided explanation towards the behavior of the model. aLIME for Anchor Local Interpretable Model-Agnostic Explanations is also an extension to LIME specifically designed to provide 'anchors' to a prediction make by the model. This framework was applied on the Part-of-Speech tagging, Image classification and Visual Question Answering models. This framework succeeds in achieving high precision and clear coverage bounds for model-agnostic models. With a view to investigate about the role of two or more attribute interactions in providing explanations for classifiers, authors Henelius *et al.* devised the ASTRID framework. Thus, it is an Automatic STRuctrue IDentification method that inspects for the largest subset of attributes and attains identical accuracy when the classifier is trained on both the original attributes and a subset of attributes. The DeepLIFT (Learning Important FeaTures) is another framework developed in 2016 by Shrikumar and Shcherbina [46]. This framework computes the score of a neuron to determine its contribution in the prediction and the explanation. The score of each neuron is computed by taking the difference between a 'reference activation' and the activation of a neuron; thus determining the deviation of the neuron; employing it on a subset dataset of the ImageNet dataset. These scores are computed using the backpropagation approach. It was later modified in 2017 [45] by the authors by employing it on the MNSIT dataset.

Recently in a 2020 publication, authors Lima and Delen performed a novel study for predicting corruption and the reason causing it across countries at various levels using modern techniques of Artificial Intelligence and Machine Learning [25]. The study was aimed at exploring the most predictive cause(s) behind corruption so that steps could be taken to eradicate them and in the long run make an attempt to make a corruption-free country wise. They collected data from 132 countries across the globe for obtaining the predictors for Corruption Prediction Index (CPI). The entire dataset was labelled into four classes. These classes were: High Corruption (Class 1),

Low Corruption (Class 2), Very High Corruption (Class 3) and Very Low Corruption (Class 4). In order to avoid biased model prediction, the k- fold ($k = 10$) cross-validation technique was applied on the dataset which splitted them into the training and testing sets. The accuracy is measured as the average of all the individual accuracy measures obtained from the k-fold cross-validation. They then employed the three popular Machine Learning approaches — Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machines (SVM). From among these models they tried to identify one model and its parameter specifications that produced unbiased and satisfactory accuracy. For comparing the three ML models, the method computed the overall accuracy as well as the class-wise accuracy. Upon experiment it was found that Random Forest was able to attain the best overall accuracy as well as class-wise accuracy among the other two models. For exploring the explainability of the prediction made by the Random forest, the variable having the most predictive power is computed using the "Actual Splitting Rate (ASR)". The ASR computes the variable importance as the ratio of the *split* factor to that of the *candidate to split* factor. This allowed for obtaining the most relevant corruption predictor. According to this ASR, the Government Integrity was splitted 90.47% of the time as compared to Property Rights, Judicial Effectiveness and Education Index which were splitted 78.84%, 77.94% and 53.93% of the times. Thus the Government Integrity is addressed as the most relevant predictor for corruption. It thus provided Government Integrity as the most relevant explanation for the corruption prediction made by the Random Forest model.

4 INTERACTIVE EXPLANATION IN MACHINE LEARNING

Wrongly identifying or predicting a concept is a severe threat to the learning models. Researchers explored ways in which the end users can contribute to such misclassifications in a number of ways. Studying the role of the end users explicitly in Machine Learning models is found to promote better learning and performance of the model. Though the involvement of the end users is negligible in developing a model for various objectives, [2] carried out an extensive study that promotes for the contribution of end users to make the Machine Learning models more interactive and understandable thus providing a relevant explanation as required by the user using it for their beneficial tasks. The authors Amershi *et al.* present different instances that highlight the significance of involving the end user in building a powerful, interactive and efficient model. Kapoor *et al.* proposed a way for interactive computation among the users and machines. They developed a system — ManiMatrix — that can effectively control the performance of the Machine Learning models in accordance with the user preference [16]. Using this application, the users can modify the decision boundary parameters in the confusion matrix via an interactive round of classification and visualization.

During the year 2011, Kulesza *et al.* proposed a *Why-oriented approach* [22] for the end users that allows them to question about a model prediction; get the explanation and later modifies the model in order for it to provide more preferable explanations in the future. The author also discusses the barriers that are faced by the end users while trying to fix the faulty behavior of the intelligent system. These barriers include Design, Selection, Coordination, Use and Understanding. The Why-oriented approach works by combining three important aspects: Firstly, it allows end users to question a prediction made by the system, secondly obtaining explanations against each prediction that provides both the current logic as well as the execution state and thirdly end users are given authority to directly manipulate the explanations to fix the system logic. All of these studies, including a few more such as [3, 51, 54] suggest that Machine Learning models and end users can work together with the aim of increasing the understandability, trust and accuracy of the model. Another idea was propounded by Kulesza *et al.* [19], where they suggested for personalizing a prediction made by the model using the *Explanatory Debugging* approach (an idea proposed by

Kulesza *et al.* in 2010 [20]). It is a task that can be performed from the users end to modify these models in an effective manner. Though a difficult task to exercise, by imbibing personalize machine learning systems into the learning model, is found to increase the understandability by 52% [21]. It is a system that explains the user about the predictions it made and the user debugs it for any correction in the learning model, thus personalizing the Machine Learning model's behavior. The user tries to identify and correct the faults in the system's reasoning that are responsible for the predictions that failed to meet the user's expectation. So the user and the learning model work as a unit and share their understanding and also influence one another. Explanatory Debugging is a controllable and satisfying approach as it focuses on users who makes the corrections by explaining predictions.

In later years in 2017, Miller *et al.* suggests that the explainable module does not take into consideration the social science viewpoint while they are developing the explainability framework which may head up towards failure in later times. Therefor one also needs to put special emphasis on how people – rather than the developers – define, generate, select, evaluate and present their explanations [34]. While trying to explain and analyze this fact, Miller *et al.* cites "beware of inmates running the asylum." The trend followed while developing the explainable AI, researchers and developers think and work by taking into consideration only their own psychological viewpoints in almost every case. But in actual scenario, it does not apply so. Since they are developing the explainable model for the audience and not for themselves, the research community should focus on people's psychology and physiological condition in general keeping in mind that the explainable model would provide these explanations to the audience. Hence the authors suggested that these explainable models should be developed in a way by incorporating social and human behavioral science in order to build an explainable AI for the audience and let them evaluate the model quality. It is so because the end users need to believe and trust these explanations provided against the predictions. This trust and belief is to a large extent dependent on the people's belief along with other psychological conditions. Thus, the author highlights that the research practitioners also should take into consideration the psychological and behavioral aspect of the people in general rather than providing explanation from the developer's perspective. Claiming whether an explanation provided is a *'good'* or *'bad'* is determined only by the researchers point of view or perception which may not always be a success in real time. Whereas it can also be viewed from from other perspectives such as —philosophy, psychology or social science. So in 2018 Tim Miller [33] suggested that the explanations provided by the XAI would prove to be beneficial only when the explanations are derived from the psychology of the audience – keeping in mind their ways of thinking, selection, presentation, etc.

5 DISCUSSION

We started our survey with an introductory note on "Explainablity" for newcomers to get idea about the very term. We provided definitions cited by a number of researchers. No doubt they are aiming at providing us with the clear concept of explainability, but one could not find an agreement on the term. It is also evident that the research community is somewhat failing in their attempt to present a satisfactory and standard level of understanding of the importance, properties, etc. in various domains. Explainability is of course a powerful tool for exploring and unboxing the working model actively involved in the prediction making. It helps the end users to take various decisions – especially that of trust in a model. Without the progress in explainability, various life changing observations could be witnessed as discussed in Section 2.1. We tried to find answer to the question in the title of the survey by walking through the approaches, tools, models developed so far in literature for explainability. The explanations provided by the various developed models and approaches are usually made with the complete knowledge of all the features. But studies and

experiments can also be carried out to investigate the influence of latent features in explanation. One of such work can be found in [23] where they worked upon latent and unobserved features. Importance and necessity of the end users in debugging the intelligent system's prediction is also an important aspect of the survey. Ways of debugging the systems have also been explored which brought into light the various challenges faced in the context. Addressing these barriers remain an unsolved research question yet to be looked upon. The role and importance of the psychological aspects of the people in proving to be beneficial for explainability models are also brought to light.

6 CONCLUSION

In this survey, we have explored the broad spectrum of explainability employed in Artificial Intelligence, Machine Learning and Deep Learning fields. Some of the noted contributions are marked and toured in this regard. Understanding explainability and its need is of course a major concern which is highlighted in this study. We presented some of the application domains that require the explainability as the actions taken after relying on the model's prediction are life changing and risky. Introspecting through this survey will guide the upcoming researchers in the field of explainability and serve as a reference in future research.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[3] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2010. Examining multiple potential models in end-user interactive concept learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1357–1360.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech [40]. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), 1–46.

[6] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803–1831.

[7] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, New York, NY, USA, 1721–1730.

[8] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, San Jose, CA, USA, 598–617.

[9] Alex A Freitas. 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter* 15, 1 (2014), 1–10.

[10] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[11] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. 2016. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974* (2016).

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[13] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[14] David J Hand. 2006. Classifier technology and the illusion of progress. *Statistical science* (2006), 1–14.

[15] Andreas Henelius, Kai Puolamäki, and Antti Ukkonen. 2017. Interpreting classifiers through attribute interactions in datasets. *arXiv preprint arXiv:1707.07576* (2017).

[16] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1343–1352.

[17] Will Knight. [n.d.]. The U.S. Military Wants Its Autonomous Machines to Explain Themselves. https://rb.gy/wtgqdi.

[18] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*. IEEE, 2556–2563.

[19] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[20] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 41–48.

[21] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10.

[22] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 1–31.

[23] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 275–284.

[24] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. 2015. Convergent Learning: Do different neural networks learn the same representations?. In *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015 (Proceedings of Machine Learning Research)*, Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar (Eds.), Vol. 44. PMLR, Montreal, Canada, 196–212.

[25] Marcio Salles Melo Lima and Dursun Delen. 2020. Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly* 37, 1 (2020), 101407.

[26] Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* 17, 4 (2001), 319–330.

[27] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. 2016. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 91–100.

[28] Stella Lowry and Gordon Macpherson. 1988. A blot on the profession. *British medical journal (Clinical research ed.)* 296, 6623 (1988), 657.

[29] Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. 2015. Deeper lstm and normalized cnn visual question answering model. *GitHub repository* 6 (2015).

[30] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.

[31] Aravindh Mahendran and Andrea Vedaldi. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5188–5196.

[32] Matt McFarland. [n.d.]. Uber shuts down self-driving operations in Arizona. https://rb.gy/vdhlwu.

[33] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[34] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[35] Saumitra Mishra, Bob L Sturm, and Simon Dixon. 2017. Local Interpretable Model-Agnostic Explanations for Music Content Analysis.. In *ISMIR*. 537–543.

[36] NewsGroup. [n.d.]. 20 newsgroups dataset for document classification. Accessed on March 2020 from http://people.csail.mit.edu/jrennie/20Newsgroups.

[37] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*. 3387–3395.

[38] Ripon Patgiri, Sabuzima Nayak, Tanya Akutota, and Bishal Paul. 2019. Machine learning: a dark side of cancer computing. *arXiv preprint arXiv:1903.07167* (2019).

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Nothing else matters: model-agnostic explanations by identifying prediction invariance. *arXiv preprint arXiv:1611.05817* (2016).

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA,

1135–1144. https://doi.org/10.1145/2939672.2939778

[41] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[43] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[44] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

[45] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3145–3153.

[46] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713* (2016).

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[48] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[49] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[50] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.

[51] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.

[52] Matt Turek. [n.d.]. Explainable Artificial Intelligence (XAI). Accessed on 20/01/2020 from https://bit.ly/38YLErX.

[53] Michael Van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 900–907.

[54] Jesse Vig, Shilad Sen, and John Riedl. 2011. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*. 93–102.

[55] Carl Vondrick, Aditya Khosla, Hamed Pirsiavash, Tomasz Malisiewicz, and Antonio Torralba. 2016. Visualizing object detection features. *International Journal of Computer Vision* 119, 2 (2016), 145–158.

[56] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[57] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*. IEEE, 2018–2025.