Engineering Applications of Artificial Intelligence 117 (2023) 105606

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

An analysis of explainability methods for convolutional neural networks

Lynn Vonder Haar, Timothy Elvira, Omar Ochoa *

Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, 1 Aerospace Blvd., Daytona Beach, FL 32114, United States of America

ARTICLE INFO

Keywords:
Explainability
Black box model
Convolutional neural network
Image recognition
High-risk fields
Safety–critical fields

ABSTRACT

Deep learning models have gained a reputation of high accuracy in many domains. Convolutional Neural Networks (CNN) are specialized towards image recognition and have high accuracy in classifying objects within images. However, CNNs are an example of a black box model, meaning that experts are unsure how they work internally to reach a classification decision. Without knowing the reasoning behind a decision, there is low confidence that CNNs will continue to make accurate decisions, so it is unsafe to use them in high-risk or safety–critical fields without first developing methods to explain their decisions. This paper is a survey and analysis of the available explainability methods for showing the reasoning behind CNN decisions.

1. Introduction

Deep learning is sophisticated in its attempts to make decisions accurately. The Convolutional Neural Network (CNN) has a reputation of encapsulating array features and excels in image recognition amongst many domains. However, like many deep learning models, a CNN does not give an explanation along with its decision, making it a black box model. Experts no longer know the rationale for its decision, resulting in a less trustworthy model due to unexplainable classifications (Burkart and Huber, 2021). Despite its accuracy for a single class, there is no assurance that the decision of the algorithm will continue to be as expected. If the network learns incorrect characteristics of an image, then it may not always correctly classify it.

While this may not be a major problem in some domains, the lack of trustworthiness is an issue in high-risk applications, such as medical diagnoses, mortgage acceptances and autonomous driving. If CNNs are making such decisions, the users must be sure that it will make the correct assessment so that they can get the proper treatment, move into a new home, or stay safe on the road (Dolph et al., 2018). In accordance with this, the European Union passed a rule that automated decisions must include a description of the reasoning behind the decision (Islam et al., 2019; Blanco-Justicia and Domingo-Ferrer, 2019).

While there is research on the black box view of many models, the specific model that this paper focuses on is CNN. This paper is a survey and analysis of available explainability methods for producing descriptions that aim to increase the trustworthiness of CNN decisions. The survey will describe deconvolutional neural networks, layer-wise relevance propagation, logic rules, inversion, Local Interpretable Model-agnostic Explanations (LIME), occlusion tests, Shapley Additive Explanations (SHAP), deep Taylor decompositions, explanation vectors, domain knowledge infusion, decision trees, and explanatory graphs as ways to explain the classification decision. This paper gives advantages and disadvantages of each of the methods, as well as a table in each section showing the domains for each of the sources used in that explanation section, giving a better understanding of the domains that explainability method best fits. The goal of this paper is to aggregate the methods so either machine learning (ML) engineers or domain experts can find corresponding domains to apply a method to their CNN model.

The layout of the rest of the paper is as follows: Section 2 discusses background of the problem, as well as an overview of CNN operations. Section 3 discusses related work in this topic. Section 4 is the survey search criteria. Section 5 is the survey of the explainability methods, including some uncommon novel methods. This section includes a description and analysis of each method and a table summarizing the domains in which that method is used. Section 6 discusses the metadata of the sources used, Section 7 gives possible future research on the topic, and Section 8 concludes the paper.

2. Background

Computer vision and CNNs are used to make predictions in many fields including medicine and banking (Islam et al., 2019; Caruana et al., 2015; Singh et al., 2020). Initially, a simple model is designed and instanced with a specific architecture. This model goes through a training process to learn important features for the classes it aims to classify. However, the trained model is unpredictable to the point where domain experts can no longer interpret the features the machine is basing its predictions on Burkart and Huber (2021). An important step in explaining the model is to understand the algorithm's domain knowledge, which is the context of the data interpretation. The domain

* Corresponding author.
E-mail address: ochoao@erau.edu (O. Ochoa).

https://doi.org/10.1016/j.engappai.2022.105606
Received 8 March 2022; Received in revised form 11 October 2022; Accepted 2 November 2022
Available online 19 November 2022
0952-1976/© 2022 Elsevier Ltd. All rights reserved.

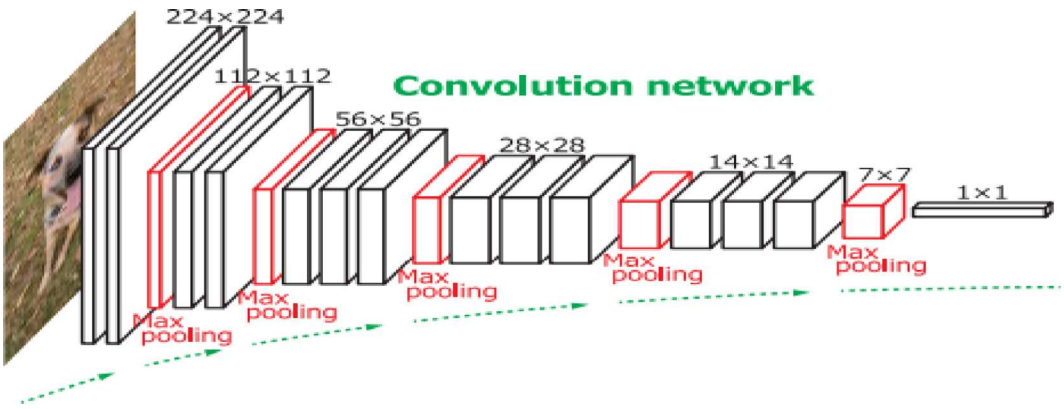


Fig. 1. Layers of a CNN. *Source:* Adapted from Noh et al. (2015).

knowledge is a set of generalized features relating to the data that can help explain the prediction simply (Islam et al., 2019). If the prediction is associated to this set of general features, then the user can begin to understand and trust the predictions made by the algorithm (Islam et al., 2019).

The primary use of CNNs is to find visual patterns among pixels for image recognition (Priyadharshini et al., 2019). The CNN is a type of artificial neural network that attempts to model the biological neural network in humans to make decisions in the same way (Sun et al., 2018; Liu, 2018b). As seen in Fig. 1, it consists of several layers of neurons with the exact number of layers depending on the model. The first few layers are convolutional layers that use neuron weights to form kernels, which are how the CNN filters the pixels (Priyadharshini et al., 2019; Liu, 2018b; Abdel-Hamid et al., 2013). These kernels aim to capture specific features in the image to form a feature map as input to the next layer (Liu, 2018b; Pinheiro and Collobert, 2014). Neuron weight represents the bias that each input into the neuron has on the output of the neuron (Sladojevic et al., 2016). Understanding what each neuron filters and activates on can increase the explainability of the network (Zhang et al., 2018a). Between each of the convolutional layers is a pooling layer that condenses the feature maps to reduce the amount of background noise passed to the next layer so that by the end of the convolutional layers, only the most important features for recognition remain (Pu et al., 2019). The pooling layers also ensure that the output of that layer is the same, even with slight changes in the input variables, which reduces overfitting (Liu, 2018b; Pinheiro and Collobert, 2014). After all of the convolutional layers are several fully connected layers that take in the feature map and use feature relevance probabilities to determine the object in the image (Pu et al., 2019). Each neuron in the last fully connected, or dense layer, represents a class of objects that the trained network recognizes (Priyadharshini et al., 2019). The importance of the convolutional layers is to reduce the number of parameters sent on to each layer. A fully connected layer requires that every neuron in the first layer connects to every neuron in the next layer, so the number of parameters for this model can grow rapidly (Albawi et al., 2017).

With the success of CNNs and, consequently, explainability methods, there is a need to qualify and quantify these explanations. Just like any human-generated explanation, machine learning explanations can do more harm than good. ML explanations may obfuscate the classification and rationale even more, further distancing the trust of the model and the user. To quantify explanations, a domain expert must examine the explanation to verify if it is properly representing the model under scope (Islam et al., 2020a,b; Rokade and Alluri, 2021). Additionally, Islam et al. provides four properties of explanation methods: expressive power, translucency, portability, and algorithmic complexity. Expressive power is the value of the structure and its ability to produce explanations. Translucency is the viability of examining an ML model in varying depths. Portability is the spread of ML models that an explainability method can cover. Algorithmic complexity corresponds to the "required time to generate an explanation" (Islam et al., 2020a,b). Lin et al. takes a more machine centric approach to quantifying explanations by valuing the explanation based on its impact on the classification. The two explainability quantifiers are decision level impact and confidence level impact which quantifies the impact of the explanation component in terms of the classification and confidence of the classification, respectively (Lin et al., 2019). Alternatively, Hoffman et al. takes a more human-centric approach that quantifies the explainability method based on user perceptions of the explanation. Hoffman et al. proposes an "explanation satisfaction scale" that a user can complete to quantify the explanation method. This scale operates similarly to a survey that has a range of questions to probe the qualitative aspects of an explanation method that a user can quantitatively grade (Hoffman et al., 2018). This approach primarily aims to quantify explainability based on user-perceived quality and may be inconsistent between users. Another quantification approach is the creation of benchmarks to test and grade explainability methods. Just like ML benchmarks aim to test the limits of an ML model, explainability benchmarks aim to probe the limitations of explainability methods. Some have studied the necessity of explainability quantification in SHAP, LIME, and saliency maps for image, text, and even audio data (Liu et al., 2021; Rokade and Alluri, 2021). Liu et al. leverages remove-and-retrain (ROAR) which gathers statistical data on a model's performance and classification tendencies each time relevant features are pruned away. Along with ROAR, Liu et al. also includes faithfulness, monotonicity, and infidelity as metrics for explanation quantification. Quantifying explainability proves difficult due to the complexity of models and the subjectiveness of the overall quality of an explainability method. Nevertheless, human-centric or machine-centric explainability quantifiers are the forefront of explainability quantification research in efforts to create a standard for ML explainability methods.

This paper explores how the CNN model learns and the methods that add a layer of explainability to better increase the trustworthiness of the architecture. There are three classes of explainability methods: gradient-based methods, perturbation-based methods, and approximation-based methods (Muddamsetty et al., 2020). Gradient-based explainability methods produce saliency maps as explanations (Muddamsetty et al., 2020). In this paper, the following gradient-based methods are discussed: deconvolutional neural netowkrs, layer-wise

Table 1
Search criteria for each section of the paper.

Section	Search criteria
Introduction and background	(Neural network AND explainability OR convolutional neural network AND explainability)
Deconvolutional neural networks	(Deconvolutional neural network AND image recognition AND explainability)
Layer-wise relevance propagation	(Layer-wise relevance propagation AND convolutional neural network AND image recognition)
Deep Taylor decomposition	(Convolutional neural network AND deep Taylor decomposition OR Machine learning AND deep Taylor decomposition)
Fuzzy logic rules	(Fuzzy logic rules AND convolutional neural network AND explainability OR logic rules AND machine learning AND image recognition)
Image inversion	(Convolutional neural network AND image recognition AND image inversion AND explainability)
Local Interpretable Model-agnostic Explanations (LIME)	(Convolutional neural network AND LIME OR Machine learning AND LIME)
Occlusion tests	(Machine learning AND occlusion tests OR Convolutional neural network AND occlusion tests OR Convolutional neural network AND perturbation explainability method)
Shapley Additive Explanations (SHAP)	(Convolutional neural network AND SHAP OR Convolutional neural network AND Shapley values OR Machine learning AND SHAP)
Explanation vectors	(Convolutional neural network AND explanation vectors AND image recognition OR machine learning AND explanation vectors AND explainability OR convolutional neural network AND quantization vectors AND explainability)
Domain knowledge infusion	(Convolutional neural network AND domain knowledge infusion AND explainability OR convolutional neural network AND domain knowledge infusion AND image recognition)
Decision trees	(Machine learning AND decision trees AND image recognition OR convolutional neural network AND decision trees)
Explanatory graphs	(Convolutional neural network AND explanatory graph AND image recognition)

relevance propagation, deep Taylor decompositions, inversion, explanation vectors, decision trees, and explanatory graphs. Perturbation-based methods changes the inputs to the model and analyzes the change in output to verify which pieces of the input were most important to the decision-making process (Muddamsetty et al., 2020). The perturbation-based methods discussed in this paper are occlusion tests. Finally, approximation-based methods also produce saliency maps, but instead of doing so by backpropagation, these methods do so via randomly checking pixel importance (Muddamsetty et al., 2020). The approximation-based methods discussed in this paper are logic rules, Local Interpretable Model-agnostic Explanations, Shapley Additive Explanations, and domain knowledge infusion.

3. Related work

Others have conducted literature surveys regarding the explainability methods of ML algorithms and specifically of CNNs. One survey describes several explainability methods currently available for Supervised Machine Learning (SML) regarding artificial neural networks, such as rule-based explanations and decision trees (Burkart and Huber, 2021). Some published papers list the domains that require explainability for decisions including medicine, justice, banking, autonomous driving, and election campaigns (Burkart and Huber, 2021). This paper focuses on CNNs specifically, rather than general SML. Additionally, this paper gives an analysis of each of the explainability methods.

Another survey goes through many of the types of explainability methods, such as feature analysis methods, saliency map methods, and explanation-by-text methods (Fan et al., 2021). However, these methods are discussed generally rather than surveying specific methods like this paper does. Additionally, the survey does not map the discussed methods to domains, which this paper does in the table following each surveyed method.

Meanwhile, another paper surveys some of the explainability methods for CNNs (Zhang and Zhu, 2018c). The survey discusses methods to disentangle the decision in the middle layers and ways to build an interpretable CNN (Andrews et al., 1995). However, this paper surveys more explainability methods, including those that do not give reasoning for middle layers. This paper also analyzes the advantages and disadvantages of each of the methods and consolidates related domains, which is novel. Such analysis suggests which explainability methods better suit certain domains, which is also not found in related work.

A different survey specifically discusses methods of extracting logic rules from artificial neural networks (Andrews et al., 1995). They follow the methods from training to usage to refining and developing a unique naming scheme for the methods discussed (Andrews et al., 1995). However, the survey describes only logic rules, whereas this paper compiles several explainability methods and introduces an analysis for each of the methods.

A survey of inversion methods in another paper splits the methods into categories based on differences in their algorithms (Jensen et al., 1999). It describes different mathematical solutions for inverting neural network decisions (Jensen et al., 1999). However, the survey does not provide analysis on the advantages and disadvantages of inversion methods, which this paper does.

Finally, another paper is a survey of explainability methods for artificial intelligence models in general and discusses some advantages and disadvantages of each (Islam et al., 2021b). However, the survey does not provide insight into the domains where each method is used and many of the methods analyzed do not overlap.

In summary, the work presented in this paper surveys several of the most popular explainability methods, specifically for CNNs. This paper presents a collection of many of the applications that use CNNs with explainability methods. It also provides an analysis of the advantages and disadvantages for each of the methods surveyed. This combination of survey and analysis can lead to the unique understanding of the explainability methods best suited for each domain.

4. Survey search criteria

This paper is a literature review of some of the more popular gradient-based, perturbation-based, and approximation-based explainability methods for CNNs, as well as some other novel methods. While this paper does not encompass all available CNN explainability methods, it provides detail on the advantages and disadvantages for each of the methods discussed and maps those methods to domains that it is commonly used in. The search engine used to find sources for this literature review was Google Scholar. The search criteria using these search engines included keywords to narrow searches to CNNs and the explainability method. Table 1 summarizes the search criteria for each section of the paper. In addition to the search criteria labeled in Table 1, the snowballing research methodology was used to find additional sources from those found with these search terms (Kitchenham and Brereton, 2013). Limitations of the research include discussions of additional variations of each of the methods presented in this paper. This is discussed further in the future work section of this paper. The inclusion criteria for this literature review were the following:

- Academic papers that discussed how the explainability methods worked.

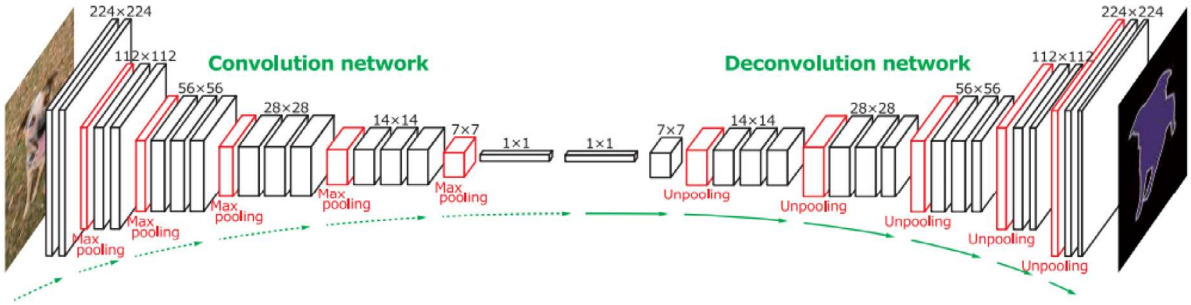


Fig. 2. Deconvolutional neural network structure.
Source: Adapted from Noh et al. (2015).

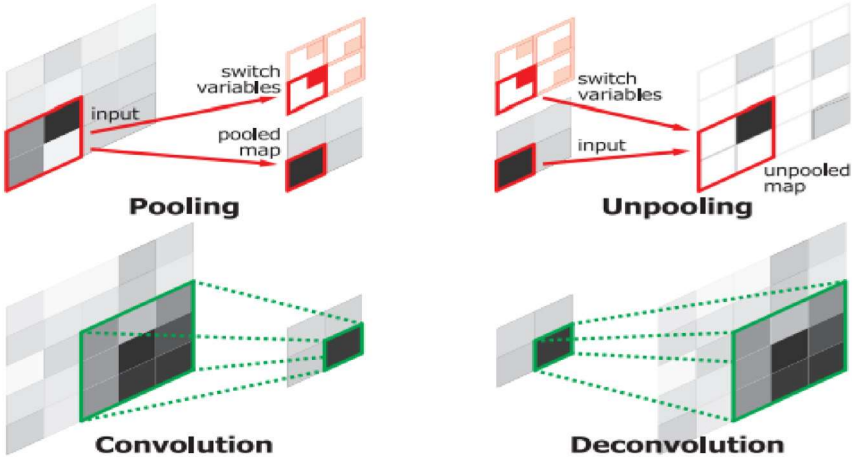


Fig. 3. Pooling versus unpooling layers.
Source: Adapted from Noh et al. (2015).

- Studies that used a CNN with an explainability method were included in the domain tables at the end of each section.
- Academic papers that described the functioning of a CNN were included in the background section.
- Surveys of available CNN explainability methods were included in the related work section.

The exclusion criteria for this literature review are as follows:

- Studies using a CNN without an explainability method.
- Academic papers focusing on datasets or model training.

5. Survey

The following section is a survey of explainability methods for CNN image recognition decisions. Each subsection includes a description of the method, followed by an analysis of the method's advantages and disadvantages and a table describing the applied domains.

5.1. Deconvolutional neural network

5.1.1. Method description

CNNs take an image and remove all unnecessary background information, using a kernel and the convolution process to find the most important features for classification. Convolutional and pooling layers are responsible for this extraction. Deconvolutional networks can take the features extracted by the CNN and use deconvolutional and unpooling layers to find the shapes of the features from the CNN and give the probability that each shape is important to classifying the image based on the neuron activations (Zeiler and Fergus, 2014; Noh et al., 2015; Mohan, 2014). Fig. 2 shows this procedure (Noh et al., 2015). The image of the dog on the far left is the input image. The CNN processes the input image and extracts the most important features. The CNN makes its prediction in the center of the figure and then the deconvolutional network on the right side of the figure unfilters and unpools the features to create a probability map as a way to explain the prediction made by the CNN (Noh et al., 2015) (Mousavi et al., 2016). The filtering layers in the CNN find the most important characteristics and filter out the rest, whereas the unfiltering layers in the deconvolutional neural network reverse the process by flipping the filters both horizontally and vertically (Mousavi et al., 2016).

The differences in the convolutional and deconvolutional network's operation are also shown in Fig. 3, where the pooling layers in the CNN reduce the image to the most important features. To form the explanation, the unpooling layers in the deconvolutional network take the reduced image and expand it to understand how the CNN made its prediction (Noh et al., 2015). The unpooling layers note the shapes that had the highest amount of activation throughout the CNN, which would give it a higher probability of usage for classification of the image (Noh et al., 2015). These activation maps are sparse since the CNN filters out many parts of the image, so the deconvolutional layers increase the density of the resulting activation maps by using mathematical operations, just as the convolutional layers do the opposite through convolutions (Zeiler and Fergus, 2014; Noh et al., 2015; Tao et al., 2017). Fig. 4 shows this where (a) is the input image, (c) is the unpooled image and (b) and (d) are deconvoluted images (Noh et al., 2015). The unpooled image shows the activated features, but with large spaces between them, whereas the deconvoluted images fill in the gaps between the activated features.

Deconvolutional neural networks directly relate to the issue discussed earlier of confidence in black box models in the medical field.

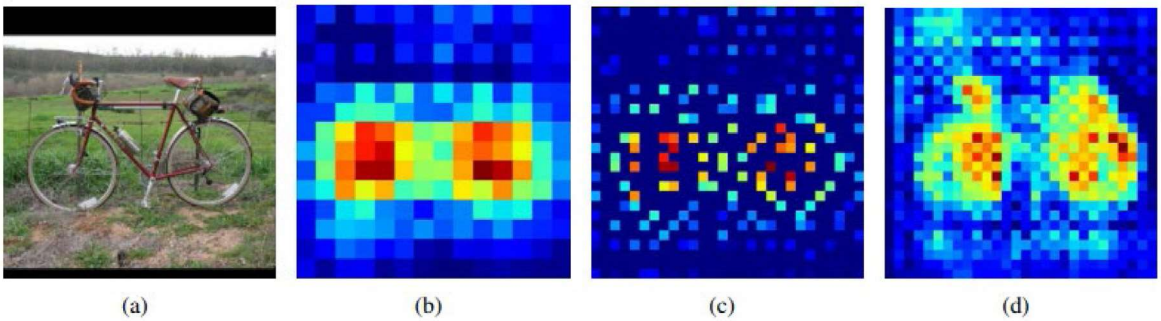


Fig. 4. Densifying the unpooled images.
Source: Adapted from Noh et al. (2015).

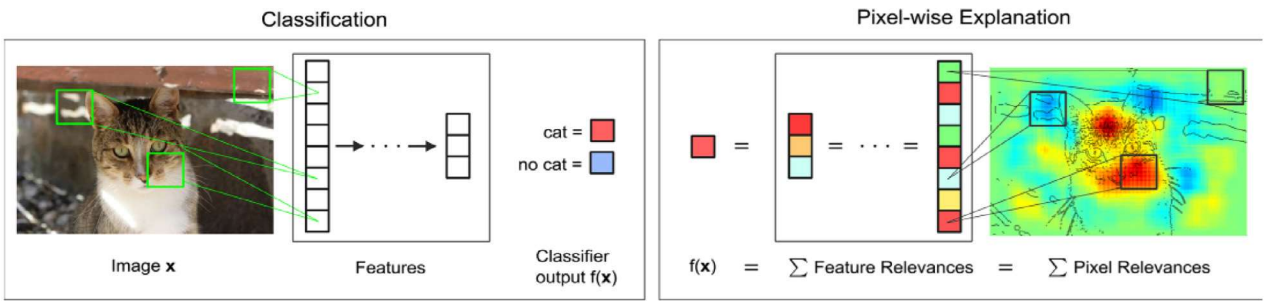


Fig. 5. Positive and negative evidence.
Source: Adapted from Bach et al. (2015).

Occasionally, a Magnetic Resonance Image (MRI) or Computed Tomography (CT) image has low resolution due to insufficient exposure time or equipment limitations (Liu et al., 2018a). CNNs can still make a prediction and diagnosis based on these images, but by adding a deconvolutional layer, the network can densify the images to increase their resolution and the readability and provide higher confidence in the diagnosis (Liu et al., 2018a; Cao et al., 2020). Convolutional and deconvolutional neural networks can also be applied together in a two-path design where the convolutional and pooling layers offer a fine-grained image based on feature activation and the deconvolutional and unpooling layers offer a course, global view of the image (Yuan, 2017). Overall, deconvolutional neural networks provide an accurate and dense activation map that represents the entire input image to explain the decision process of the CNN.

5.1.2. Method analysis

Deconvolutional neural networks have many advantages, including that they are relatively simple to understand given an understanding of CNNs because they perform opposite functions to the CNN (Noh et al., 2015). For example, knowledge of the pooling layers will help in understanding unpooling layers. Additionally, deconvolutional neural networks give a visual representation of the decision explanation in the form of a probability map and an activation map (Noh et al., 2015). Finally, and most importantly, deconvolutional neural networks densify their explanation (Noh et al., 2015). While other explainability methods may just show a map of the important regions left at the end, deconvolutional neural networks densify the image while retracing the neural activations. The result is an activation map much closer to the original image.

However, there are some drawbacks to using this explainability method. Primarily, the deconvolutional neural network can repeat analysis of some pixels, causing a checkerboard effect that changes the appearance of the reconstructed image (Gao et al., 2017; Odena et al., 2016). Additionally, as seen in Fig. 4, even with densification, it could be difficult to see the original image in the reconstruction without the proper architecture. Therefore, even with densification of the reconstructed image, the explanation could be difficult to understand and verify. Finally, deconvolution only shows the relative contribution of each region to the decision. In Fig. 4, the blue areas were low relative contribution while yellow and red are higher relative contribution. This explanation does not distinguish the regions that suggested a different decision.

5.1.3. Applications

Table 2 summarizes the sources used in this section on deconvolutional neural networks. The summary includes the domain of the source and the application of the model.

5.2. Layer-wise relevance propagation

5.2.1. Method description

The goal of layer-wise relevance propagation (LWRP) is to understand how much each pixel contributes to the decision (Bach et al., 2015). Through pixel-wise decomposition, the probability that each pixel contributes to the decision is determined (Bach et al., 2015). Pixels depicting the background will not have contributed much to the decision, but pixels depicting distinct aspects of the object will contribute to the decision. Different from many other methods of CNN explainability, LWRP not only produces a heat map of pixel contribution, but also which pixels suggested a decision different from the one made, which is called positive and negative evidence, respectively (Samek et al., 2017a; Grezmak et al., 2020). This is shown in Fig. 5 where the blue spots in the pixel-wise explanation do not indicate that a cat is in the picture, but the red spots do (Bach et al., 2015; Zintgraf et al., 2017). This increases the confidence in the decision made by the CNN and increases the precision of the explanation because it is not only considering the helpful pixels, but also the usually disregarded or useless background pixels.

LWRP builds on the idea that a neural network reaches its decision through several layers, rather than a single computational layer (Bach et al., 2015). It combines the relevance of each output pixel to build

Table 2
Domain summary for deconvolutional neural networks.

Title and reference	Domain	Application
Visualizing and Understanding Convolutional Networks (Zeiler and Fergus, 2014)	Image recognition	Using visualizations to find problems with the model
Learning Deconvolution Network for Semantic Segmentation (Noh et al., 2015)	Semantic segmentation algorithm	Using deconvolutional neural networks to create activation maps for convolutional neural networks
Deep Deconvolutional Networks for Scene Parsing (Mohan, 2014)	Scene Parsing	Recognizing objects within an image by classifying each individual pixel in the image
Understanding How Deep Neural Networks Learn Face Expressions (Mousavi et al., 2016)	Facial expression recognition	Training a CNN on emotional expressions and using a deconvolutional neural network to understand the learning process
Unsupervised-Restricted Deconvolutional Neural Network for Very High-Resolution Remote-Sensing Image Classification (Tao et al., 2017)	Very high-resolution satellite imaging	Earth observation for applications such as land-cover mapping, road extraction, and object recognition
Learning Deconvolutional Deep Neural Network for High Resolution Medical Image Reconstruction (Liu et al., 2018a)	Medical diagnoses	Deconvolutional neural networks can densify low resolution scans to make medical diagnoses easier and more confident
Deconvolutional Neural Network for Image Super-Resolution (Cao et al., 2020)	Single image super-resolution	Producing a high-resolution image from a low-resolution image
Hierarchical Convolutional-Deconvolutional Neural Networks for Automatic Liver and Tumor Segmentation (Yuan, 2017)	Medical diagnosis	Combining convolutional and deconvolutional methods to have both a local and global view of the decision process

a relevance map for that layer, then continues backwards towards the input image to determine the extent that each pixel contributed to the decision (Grezmak et al., 2020; Samek et al., 2017b; Grezmak et al., 2019). This method works backwards from the prediction made by the CNN and uses the weights and activations of the neurons to determine the positive and negative evidence values of the pixels both connected to and surrounding neurons (Samek et al., 2017a,b). As shown in Eq. (1), the sum of the relevance for each pixel is constant between layers (Grezmak et al., 2019). This rule ensures that the explanation stays consistent with the decision throughout the propagation (Samek et al., 2017a). In this way, it is possible to gain an understanding of what each layer filtered out or found important, which is demonstrated in Fig. 6 (Grezmak et al., 2019). Fig. 6 shows the classification process using the CNN at the top of the figure and the LWRP explanation process at the bottom. The LWRP processes the decision that there is a missing tooth in the gear and the relevance of each pixel in the final layer of the CNN. It then propagates the relevance backwards and populates the heat map accordingly resulting in a concise heat map of the most important features.

$$f(x) = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \sum_{d} R_d^{(l)}$$

Equation 1. Sum of relevance equation. Source: Adapted from (Grezmak et al., 2019).

The weights of the neurons in each layer indicate what part of the image was filtered out by that layer, reducing the probability of that region affecting the decision (Samek et al., 2017a). LWRP is set apart from other methods of CNN explainability by using both positive and negative evidence to be more precise about the decision made. While other methods either ignore the parts of the image that did not support the decision, or just give them a low probability, LWRP shows that those regions, such as background regions, opposing the decision made by the CNN.

5.2.2. Method analysis

The main advantage of LWRP is that it provides both positive and negative evidence (Bach et al., 2015) . This gives a more detailed explanation of the decision because rather than showing the relative contributions of each region, LWRP shows the regions that supported the decision made, and the regions that supported a different decision. This could work well towards increasing confidence in the model's accuracy because the explanation gives more of a big picture perspective. Additionally, the model analyzes the role of individual pixels in the decision-making process, so the explanation should be precise in exactly which regions supported the decision and which regions did not (Grezmak et al., 2020). Finally, the sum of the relevance scores in each layer stays constant, which means that each layer has as much of an impact on the decision as all the others (Grezmak et al., 2019). This should help to dissipate biases during the propagation of the explanation.

A main disadvantage of LWRP is that if the method hits a dead end in relevance analysis, it may just disperse the relevance to surrounding pixels (Bach et al., 2015). Another disadvantage to analyzing the relevance of each individual pixel could be that the explanation depends on such a small region that there could be higher overlapping of potential decisions. A single pixel could indicate multiple possible decisions, so it would be difficult to tell if the explanation is correct. With the relevance score of each pixel affecting the scores of its neighboring pixels, this could have a snowball effect.

5.2.3. Applications

Table 3 summarizes the sources used in this section on LWRP. The summary includes the domain of the source and the application of the model.

5.3. Deep Taylor decomposition

5.3.1. Method description

Deep Taylor decomposition is a white box explainability method based on the understanding that a CNN decision is composed of more than one computation (Montavon et al., 2017). In this way, the final decision is decomposed through backpropagation and Taylor expansions on each neuron (Montavon et al., 2017; Kauffmann et al., 2020;

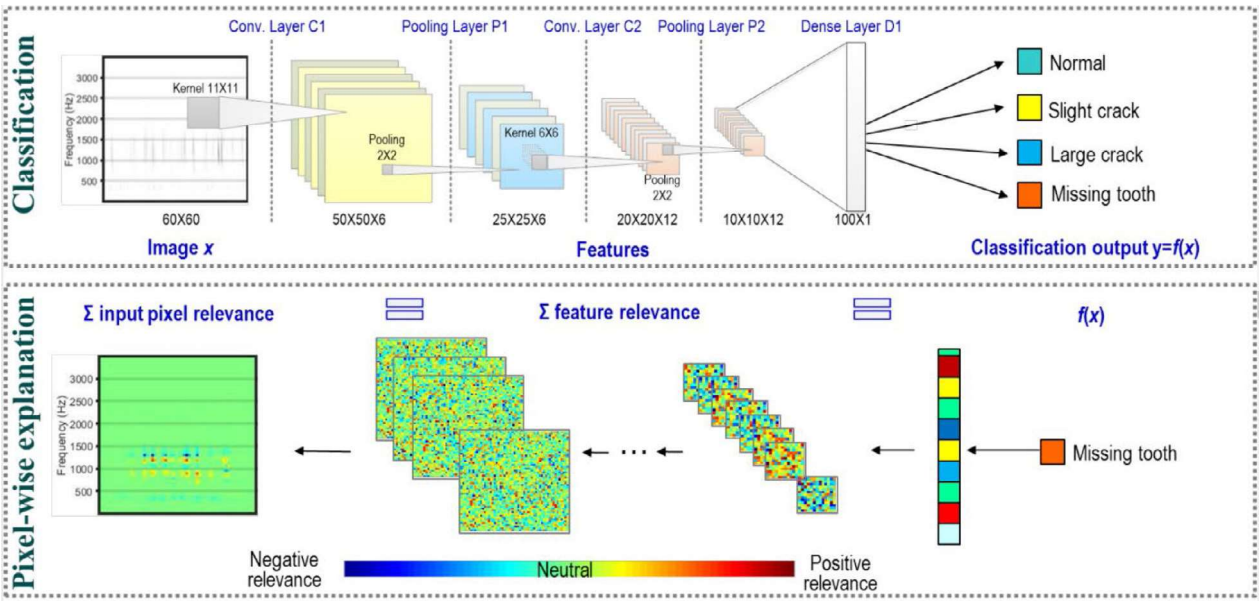


Fig. 6. LWRP explanation.
Source: Adapted from Grezmak et al. (2019).

Table 3
Domain summary for LWRP.

Title and reference	Domain	Application
On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation (Bach et al., 2015)	Image recognition	Using LWRP to explain decisions for both CNNs and Bag of Words
Evaluating the Visualization of What a Deep Neural Network Has Learned (Samek et al., 2017a)	Image recognition	Using multiple methods (sensitivity analysis, deconvolution, and LWRP) to create heat maps describing the same image recognition decision
Interpretable Convolutional Neural Network Through Layer-Wise Relevance Propagation for Machine Fault Diagnosis (Grezmak et al., 2020)	Machinery condition monitoring	Using CNNs to diagnose faults in machinery. The training for the CNN uses fault patterns to develop the diagnostic algorithm
Visualizing Deep Neural Network Decisions: Prediction Difference Analysis (Zintgraf et al., 2017)	Image recognition	Using positive and negative evidence to characterize the object in the image. This method is helpful in healthcare diagnoses
Explainable Artificial Intelligence: Understanding, Visualizing, and Interpreting Deep Learning Models (Samek et al., 2017b)	Explaining deep learning decisions	Application examples include board game strategy, object detection and natural language processing
Explainable Convolutional Neural Network for Gearbox Fault Diagnosis (Grezmak et al., 2019)	Analyzing mechanical failure	Using image recognition to find the reason for faults in gearboxes

Kindermans et al., 2016; Bazen and Joutard, 2013). The explanation starts at the output and contributes relevance to each activated neuron in that layer before backwards propagating and doing the same function on the previous layer (Flutura et al., 2020). This computation continues until the model reaches the input image and relevance values are assigned to each pixel of the input image (Flutura et al., 2020). This process results in a saliency map showing the contribution of each pixel to the final decision, as shown in Fig. 7 (Flutura et al., 2020). This process is similar to LWRP, but with the deep Taylor method, the decomposition in each layer is a simpler computation (Montavon et al., 2019). Just like with LWRP, deep Taylor decomposition requires that the total relevance of each layer is constant (Kindermans et al., 2016).

As seen in Fig. 7, the result of deep Taylor decomposition is the contribution of each pixel to the final decision (Flutura et al., 2020). In the figure, the dark pixels represent the background pixels that did not contribute to the decision. This saliency map is simpler than that shown in Fig. 5, which illustrates the positive and negative evidence from LWRP. However, both methods have advantages based on the requirements of the domain and application.

5.3.2. Method analysis

An advantage that deep Taylor decomposition has over its counterpart LWRP, is that the computations are simpler. However, this introduces more instability in the explanation, making its contribution to user confidence less than that of LWRP (Montavon et al., 2019). This is especially true for more complex models, such as deep neural networks (Montavon et al., 2019). Additionally, deep Taylor decomposition results in simpler saliency maps than LWRP, as seen in Fig. 7. This saliency map does not show the negative evidence of the pixels that did not contribute to the decision, but it does highlight the pixels that did contribute.

Table 4
Domain summary for deep Taylor decomposition.

Title and reference	Domain	Application
Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition (Montavon et al., 2017)	Image recognition	Applying explainability methods to CNN image recognition
Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models (Kauffmann et al., 2020)	Image recognition	Applying the deep Taylor decomposition method to single-class models
Interactive Machine Learning and Explainability in Mobile Classification of Forest Aesthetics (Flutura et al., 2020)	Image recognition	Tree and forest characteristic classification
Layer-Wise Relevance Propagation: An Overview (Montavon et al., 2019)	Explainability of deep neural network decisions	Introducing LWRP and comparing it to the simpler deep Taylor decomposition method
Investigating the Influence of Noise and Distractors on the Interpretation of Neural Networks (Kindermans et al., 2016)	Neuroimaging	Analyzing explainability methods robust against noise and distractors for use in neuroimaging
The Taylor Decomposition: A Unified Generalization of the Oaxaca Method to Nonlinear Models (Bazen and Joutard, 2013)	Image recognition	Increasing utilization of he Oaxaca decomposition method

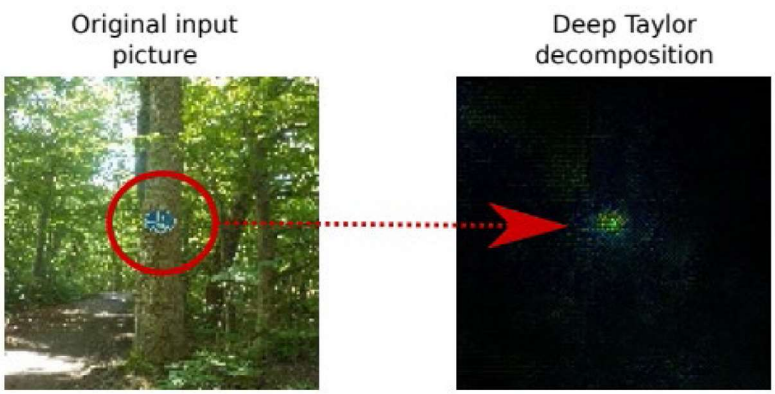


Fig. 7. Saliency map from deep Taylor decomposition.
Source: Adapted from Flutura et al. (2020).

$$e = \langle r = \{age \leq 25, job = none, amount > 5k\} \rightarrow deny,$$
$$\Phi = \{(\{age > 25, amount \leq 5k\} \rightarrow grant),$$
$$(\{job = clerk, car = yes\} \rightarrow grant)\}\rangle$$

Fig. 8. An example of decision and counterfactual rules.
Source: Adapted from Guidotti et al. (2018).

5.3.3. Applications

Table 4 summarizes the sources used in this section on deep Taylor decomposition. The summary includes the domain of the source and the application of the model.

5.4. Logic rules

5.4.1. Method description

Fuzzy logic rules can also explain the decisions made by CNNs. While many explanatory methods for CNNs produce some sort of image, logic rules describe the reasoning in sentences (Bologna, 2019). These logic rules provide a much more detailed description of the decisions than methods that highlight the important sub regions of the image (Bologna, 2019). Their main format is an if-then statement to form a linguistic explanation of the decision (Usman et al., 2021). These if-then rules can make logical inferences in the future (Andrews et al., 1995; Zadeh, 1996). The writing of the if-then statements occurs from the initial data set. Modifications to the if-then statements can occur during the CNN training process. Additionally, modifications can occur from both objective and subjective data as an if-then rule (Bonanno et al., 2017).

The fuzzy logic rules can be either ordered, such as with an if-else statement, or they can be unordered, where each rule is its own separate if statement (Bologna, 2019). When considering a local explanation, there are two different types of rules. A logic rule is a straightforward explanation of why the CNN labeled a sub region of the image or filter (Guidotti et al., 2018). A counterfactual rule shows how to change the conditions of the sub region or filter to invert it (Guidotti et al., 2018). Combining both types of rules can form a decision rule, which is a local explanation of the decision made (Guidotti et al., 2018). The local explanations feed into the final decision and backwards propagate for a better understanding of the decision. For example, Fig. 8 shows a decision rule e for denying a loan. The two following rules, denoted by Φ are counterfactual rules that would invert the decision of the black box model (Guidotti et al., 2018). As shown, this explanation method relies on mathematical logic. Writing the rules as if-then statements or restructuring them into a decision tree, facilitates interpretability by humans.

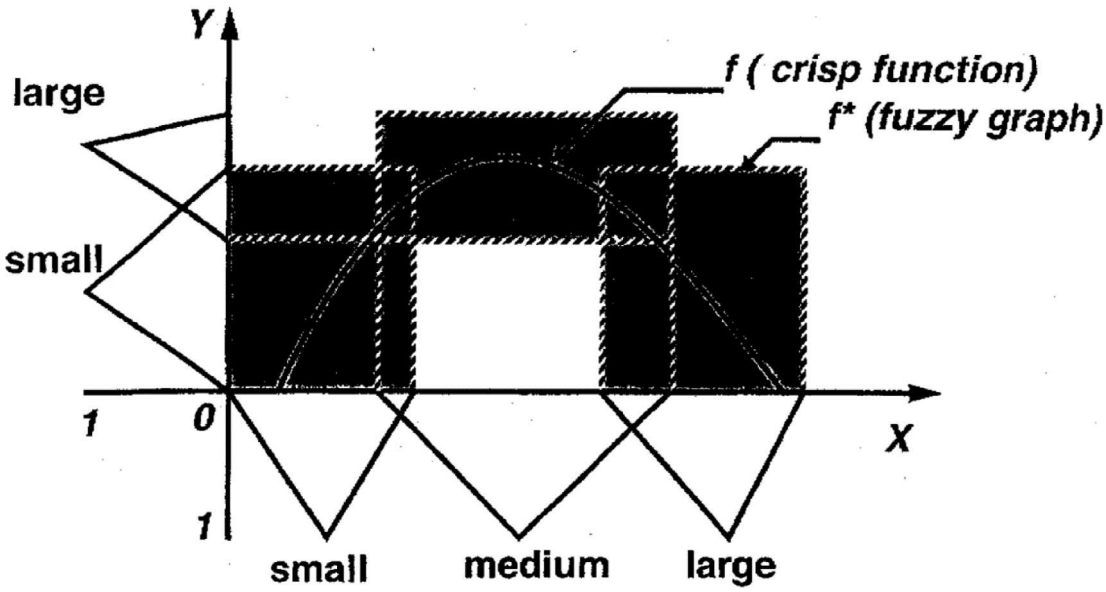


Fig. 9. The actual function versus the fuzzy approximation.
Source: Adapted from Zadeh (1996).

A description of fuzzy logic is computing with words rather than numbers and symbols (Zadeh, 1996). Fuzzy logic groups similar elements together to form a granule. Keywords describe granules (Zadeh, 1996). When making decisions, such as in image classification, the model can granulize elements of the image to get a pool of keywords as an explanation of the decision. This explanation includes mathematical representations in the form of fuzzy graphs. These fuzzy graphs approximate the real function. Fig. 9 shows the actual function as the line while the fuzzy approximation of the function are the boxes surrounding it (Zadeh, 1996). In image classification, functions can represent images with each surrounding box being a related granule that explains the reasoning behind the decision.

The structure of the CNN can accommodate fuzzy logic by adding a fuzzy logic layer. Generally, the addition of the fuzzy logic layer is near the fully-connected layer, or it replaces the fully-connected layer (Bologna, 2019; Xi and Panoutsos, 2018). With the extra layer being near the final classification of the image, it can extract the logic rules from the classification layer (Xi and Panoutsos, 2018). The logic layer backwards propagates the rules it forms to the input layer to find the set of words used to describe the decision-making process (Bologna, 2019). The activation of certain filters in each layer contributes words to the end propositional rule, so backtracking with the words can show which filters activate in each layer to indicate what the neural network found important to classifying the image.

5.4.2. Method analysis

Logic rules describe not only the region of importance for the decision, but also why that region affected the decision. Their descriptiveness makes it even easier to understand the explanation. It can also be easier to see changes in the model by viewing changes in the if-statements. With other explainability methods, there could be a slight change in regions, but the changes could be more difficult to see than with logic rules.

This method should be simple to understand because words describe the rules, which makes the explanation good for a non-expert in the field. The reasons behind the decision made for image recognition, however, could be more complex than can easily be expressed in a series of simple if-statements. When explaining convolution, complex mathematics and concepts could bog down the descriptions of the if-statements. While a mathematician might be able to understand the rules, a non-expert, or even an expert in a non-mathematical field, for example a doctor, might not understand the explanation, thereby vastly decreasing its use.

5.4.3. Applications

Table 5 summarizes the sources used in this section on logic rules. The summary includes the domain of the source and the application of the model.

5.5. Inversion

5.5.1. Method description

Inversion of a CNN takes a convoluted image and attempts to reconstruct the original image. Each layer of the network filters out part of the image that it does not deem necessary for categorization, so inverting the image after each layer can indicate what the CNN felt was important to image recognition and how it made its decision (Mahendran and Vedaldi, 2015). This can help to understand the neural weights in each layer of the network to indicate what characteristics each layer filters.

There are two methods to inversion. The first is optimization-based inversion which uses gradients to invert the image after each convoluted layer, which requires a white box model for gradient computation (Yang et al., 2019). The other approach is training-based inversion where the machine uses a second model, called the inversion model, to invert the image to see the filtering of the convolutional layer (Yang et al., 2019). Inversion after the convolutional layers still largely looks like the original image, so the network is working to filter the background and narrow its focus to the main object of the image, as shown in Fig. 10 (Mahendran and Vedaldi, 2015). There are different methods of inversion based on the mathematical model used, but despite the differences in models, the reconstructed image is equivalent.

Functions represent convoluted images and can be mathematically inverted. The original representation of the image and the inversion algorithm used changes how the reconstructed image will look (Mahendran and Vedaldi, 2015). However, the general outline of the main building is still visible in all of the reconstructed images in Fig. 10, so in all cases, it is clear which pieces of the image the CNN found important for image classification. Specifics of how a function is modeled leads to the differences in the reconstructed images. As with any function that is indefinitely integrated, there are a family of solutions, shown in Fig. 11 (Jensen et al., 1999). Just as there are an infinite number of constant values, there are an infinite number of possible reconstructed images (Jensen et al., 1999). This causes the differences in the reconstruction of the image in Fig. 10.

Table 5
Domain summary for logic rules.

Title and reference	Domain	Application
Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks (Andrews et al., 1995)	Artificial neural network decision processes	A survey reviewing the current techniques for producing fuzzy logic rules from inputs and outputs of artificial neural networks
A Simple Convolutional Neural Network with Rule Extraction (Bologna, 2019)	Discretized Interpretable Multi-Layer Perceptron (DIMLP)	The main use of DIMLP is to explain decisions made with textual data. It forms propositional logic rules
A Human-in-the-Loop Probabilistic CNN-Fuzzy Logic Framework for Accident Prediction in Vehicular Networks (Usman et al., 2021)	Automated driving	More accurate accident prediction considering the emotions of drivers through use of a CNN
Fuzzy Logic = Computing with Words (Zadeh, 1996)	Computing with words	Similar elements are grouped together to form a granule with each granule being represented by a keyword. Decisions can be backtracked to find these keywords
An Approach to Explainable Deep Learning Using Fuzzy Inference (Bonanno et al., 2017)	Explainability of CNN decisions	Creating fuzzy inference by adding fuzzy rules to the deep learning model
Local Rule-Based Explanations of Black-Box Decision Systems (Guidotti et al., 2018)	Decision tree local classifier built by a genetic algorithm	Using genetic algorithms to form rules that are then stored in a decision tree. This study focuses on local explanations which explains specific decisions, rather than an explanation of how the entire model functions
Interpretable Machine Learning: Convolutional Neural Networks with RBF Fuzzy Logic Classification Rules (Xi and Panoutsos, 2018)	Radial-Basis-Function Neural-Fuzzy Layer	Adding a fuzzy logic layer to the CNN to extract the logic rules while the neural network is making its decision

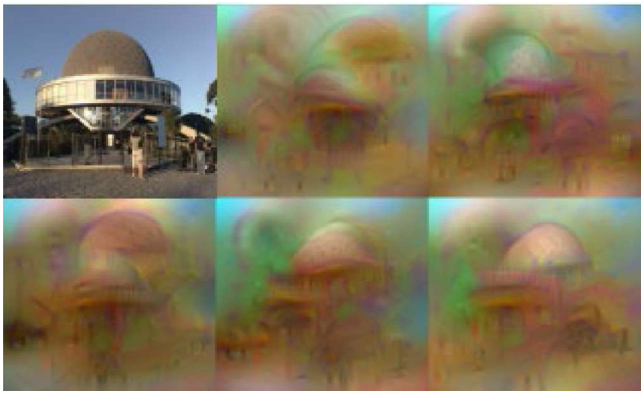


Fig. 10. Five inversion variations of the original image in the top left corner.
Source: Adapted from Mahendran and Vedaldi (2015).

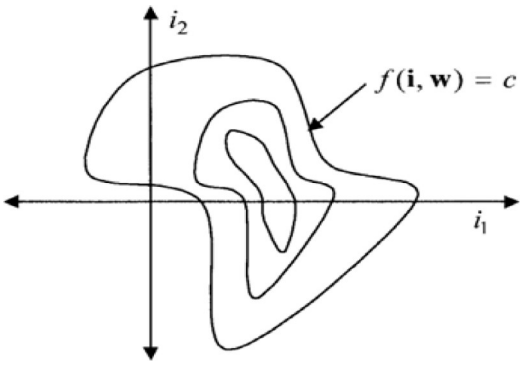


Fig. 11. A family of solutions of an original function. Represents how there are many ways to invert the original image.
Source: Adapted from Jensen et al. (1999).

Fig. 12 shows the reconstruction of the image after each layer (Mahendran and Vedaldi, 2015). The top left image is the pooled image after the first convolutional layer. Each consecutive image is the result of the next layer in the network. After each layer, the reconstructed image looks a little less like the original image, but it shows the most important parts used for classification (Mahendran and Vedaldi, 2015). In this case, the face is the most important feature. Fig. 12 is a detailed representation of the main goal of neural network inversion. By inverting the image after each layer in the CNN, it is possible to understand what each layer found important to recognizing the object in the image. An extension of that can apply to finding the relative neural weights of each layer.

5.5.2. Method analysis

Identifying the regions of the image filtered out in each convolutional layer should be easier using inversion because it reconstructs the image for comparison to the original, making the decision process simple to follow. The early layers begin to filter out the background regions and the later layers only keep the most important outlines of figures in the image. Additionally, the mathematics behind inversion is simple once the computer has the function representing the image. This makes the explanation simple for non-mathematical experts to understand.

However, a drawback to this is that there is a whole family of functions that could represent the image. This means that when replicating the decisions, if the function is different enough the next time the algorithm runs, the explanation could be significantly different. The irreproducibility of the explanation could have major negative effects on the user's confidence. Additionally, inversion introduces an optimization problem connected to switching the direction of computation and making the outputs into inputs (Mahendran and Vedaldi, 2015) .

5.5.3. Applications

Table 6 summarizes the sources used in this section on inversion. The summary includes the domain of the source and the application of the model.

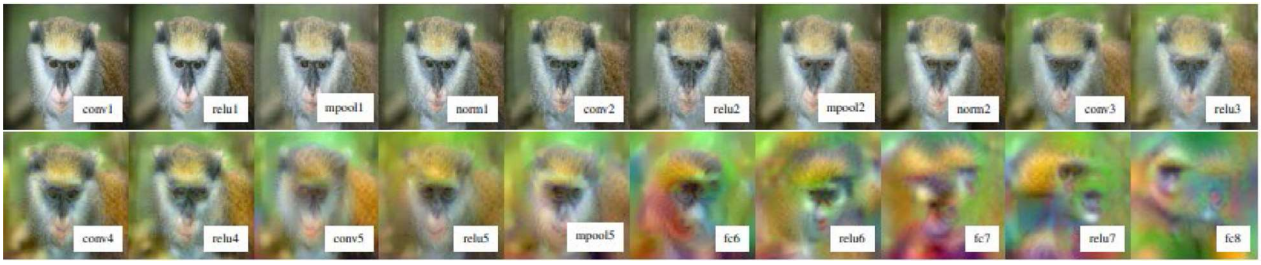


Fig. 12. Inversion of the image after each layer of the CNN.
Source: Adapted from Mahendran and Vedaldi (2015).

Table 6
Domain summary for inversion.

Title and reference	Domain	Application
Inversion of Feedforward Neural Networks: Algorithms and Applications (Jensen et al., 1999)	Multi-element evolutionary inversion, constrained neural network inversion	Query-based learning, sonar performance analysis, power system security assessment
Understanding Deep Image Representations by Inverting Them (Mahendran and Vedaldi, 2015)	Histogram of Oriented Gradients (HOG)	Explainability of black box image classification using gradient optimization
Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment (Yang et al., 2019)	Training-based model inversion	Adversarial settings
Towards Explanation of DNN-Based Prediction with Guided Feature Inversion (Du et al., 2018)	Instance-level deep neural network interpretation	Producing saliency maps from images using deep neural networks

5.6. Local Interpretable Model-agnostic Explanations (LIME)

5.6.1. Method description

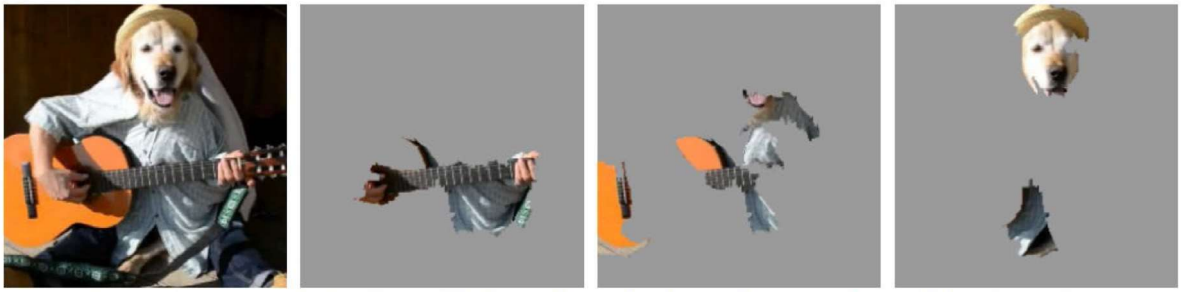
LIME, or Local Interpretable Model-agnostic Explanations, is a black box approach that provides interpretability to a model's classification regardless of the original model architecture. LIME's agnostic nature allows the method to be paired with any classifier trained on any form of input data by treating the original model as a black box (Ribeiro et al., 2016). This approach aims to bring interpretability for otherwise low interpretable models up to par with naturally interpretable models (Ribeiro et al., 2016). LIME approximates the classification of a model within another interpretable model. Typically, LIME's explanations are patches in an image or words in text that correspond to its classification. Ribeiro et al. conduct two experiments, image classification with neural networks and text classification with random forests, using LIME to show the approach's modularity to both model type and classification problem (Ribeiro et al., 2016). LIME aims to reduce the locality-aware loss to ensure that the explanation is "locally-faithful" rather than globally faithful to the classifier (Ribeiro et al., 2016). The locality-aware loss is composed of the set of potential interpretable models, interpretable components of the classification, and the complexity of the explanation. LIME reduces the locality-aware loss while also maintaining low complexity of the explanation to maintain human-understandability (Ribeiro et al., 2016). For text explanations, a bag of words is used to locally explain a corpus for a human-in-the-loop to consider when weighing a classifier's inference. For images, LIME uses "super-pixels" which are binary pixels that are either apparent or pruned away depending on their pertinence to the classification. Ultimately, this explainability method aims to garner trust between the model and the human-in-the-loop user (see Fig. 13).

LIME is being extended and altered to fit many new application spaces and domains. Mishra et al. leverage LIME for music content analysis, termed S-LIME, or Sound LIME. S-LIME is extended to time series deep learning models by bringing explanations to the time or range that is most pertinent to a classification. Mishra et al. proposes that this adds a "temporal and spectral" annotation when explaining a model inference (Mishra et al., 2017).

Deterministic Local Interpretable Model-Agnostic Explanations, conceived and developed by Zafar and Khan, or also known as DLIME is an iteration of LIME that is altered to be more deterministic in its explanations. Due to LIME relying on non-deterministically perturbing points, like when used in images and "super-pixels", there is inherent instability in the explanation method (Zafar and Khan, 2019). Zafar and Khan propose using hierarchical clustering to use as the interpretable model rather non-deterministically perturbing data points. The authors' use case and experiment applies DLIME to computer-aided diagnostic systems. The authors find that DLIME provides stable and consistent explanations for similar datasets compared to LIME generating different explanations due to its inherent non-determinism.

Hakkuom et al. applies LIME to the medical field by using LIME to generate explanations to two deep learning models, MLP and Deep MLP, trained on breast cancer datasets. Hakkuom et al. claims that their generated explanations helped provide insight on the model's behavior which, in turn, elevated the trustworthiness of the cancer classification and diagnosis (Hakkoum et al., 2020).

Visani et al. proposes an optimized alternative to LIME, coined OptiLIME. Interestingly, Visani et al. also note LIME's instability similarly to Hakkuom et al.; however, the authors of OptiLIME believe it stems from the inconsistent mathematical conditions that the explanations are generated under. Furthermore, Visani et al. note that the only means of ensuring stability is to enforce explanation generation under the same conditions to ensure statical equality (Visani et al., 2020). OptiLIME uses an adherence metric to force condition similarity thus incorporating stability into LIME. The authors use a benchmark data set focused on medical research to test OptiLIME. The author's main findings are examining the trade-off space between condition adherence and stability allowing for users to select their preferred explanation stability range. The authors found further interesting details when examining explanation conditions within LIME like a simple linear regression serves as the best interpretable model or ridge regularization is not required.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Fig. 13. LIME's explanation of an image using super-pixels to show locally related features pertinent to the classification.
Source: Adapted from Ribeiro et al. (2016).

Another method, E-LIME, combines LIME with genetic algorithms to reduce the required processing power for LIME computations and to introduce variability into the number of sampled superpixels (Wang et al., 2021). The genetic algorithm makes choosing the samples of perturbed images for the most accurate explanation much faster and computationally efficient (Wang et al., 2021). Such a combination makes LIME feasible to use in more domains and applications.

5.6.2. Method analysis

Overall, LIME is a useful black box explainability model that locally maps a classification using other interpretable models. While this method ensures a black box view, there are tradeoffs like its non-determinism and instability to generate the same explanations for a given classification. The use of more interpretable machine learning algorithms to explain the results of a less-interpretable model compounds the non-determinism. Additionally, there is no guarantee that both models learned the problem the same way or that if either model, regardless of interpretability, can learn the complexities of the problem. There is no definitive way to prove this either, as LIME is a black box approach in which either models cannot have a white box view or tests. The black box approach can also be a valuable trait as it does not require in-depth programming or mathematics to dissect the white box view of a more complex learning algorithm. Furthermore, this agnostic quality allows LIME to be extensible to many models and domains. While many current research efforts are pushing LIME to explore more domains, a heavy focus is improving LIME to shore up its instability and non-deterministic deficiencies. One specific domain where LIME is heavily prevalent is the medical diagnosis explanation domain (Zafar and Khan, 2019; Togacar et al., 2022; Palatnik de Sousa et al., 2019). The predominant use of LIME in medical diagnosis domains could be largely due to the use of images being heavily used in the diagnosis process. While ML medical diagnosis is still largely a human-in-the-loop process, LIME adds an additional layer of trust between a physician and an ML prediction. While LIME's non-determinism is a limiting factor, its versatility in a plethora of learning problems proves LIME's usefulness in being a proficient explanation method for more complex models.

5.6.3. Applications

Table 7 summarizes the sources used in this section on LIME. The summary includes the domain of the source and the application of the model.

5.7. Occlusion tests

5.7.1. Method description

Occlusion tests are black box perturbation-based methods of CNN explainability (Muddamsetty et al., 2020). They work by obscuring or changing sections of an image and analyzing how the prediction or model confidence in the decision changes to determine which parts of the image had the largest effect on the output without knowing how the layers of the model work (Muddamsetty et al., 2020; Petsiuk et al., 2018; Hemelings et al., 2021; Ieracitano et al., 2021a,b). Based on the information gleaned from each occlusion, the saliency is mapped as a visual explanation of the decision-making process (Nigri et al., 2020; Ieracitano et al., 2021a,b). Occlusion can be performed by setting the value of each pixel in a certain region to the same value (Nigri et al., 2020). Additionally, parts of an image can be overlapped by another image to hide pieces important for classification, as shown in Fig. 14 (Osherov and Lindenbaum, 2017).

Occlusion is especially important to measuring the accuracy of a CNN decision because studies have shown that partial occlusion of an image can lead to reduced accuracy of the decision (Osherov and Lindenbaum, 2017; Nigri et al., 2020). This is especially true if the model is trained on completely unoccluded training instances (Osherov and Lindenbaum, 2017). Therefore, occlusion is not only used for CNN explainability, but also for improving CNN accuracy (Osherov and Lindenbaum, 2017). By training on partially occluded data instances, the model becomes adept at handling instances where the main object may not be completely visible (Osherov and Lindenbaum, 2017). Therefore, occlusion may increase user confidence in a CNN not only by producing a saliency map explaining the decision-making process, but also by training the model on a more challenging dataset.

Occlusion test is an umbrella term for many variations of testing. However, each variation follows the same pattern of hiding parts of the image and analyzing the change in the model output. Examples of specific occlusion methods include Randomized Input Sampling for Explanation (RISE) and Time Series Randomized Evolving Input Sample Explanation (TimeREISE) (Petsiuk et al., 2018; Mercier et al., 2022). RISE is an occlusion method that randomly masks parts of the input image to build a saliency map (Petsiuk et al., 2018). RISE sub-samples each image to create a whole dataset to evaluate and the resulting saliency map reflects the weights from each of the sub-sampled images (Petsiuk et al., 2018). In contrast, TimeREISE uses time series classification to explain decisions (Mercier et al., 2022). TimeREISE does use random masks similar to RISE, but it masks over time, as well as over attributes so that it can be applied to time series data (Mercier et al., 2022). This method also changes the occlusion density so that the number of relevant data points for explanation can change (Mercier et al., 2022). In addition, the Similarity Difference and Uniqueness (SIDU) method falls under the umbrella of occlusion tests (Muddamsetty et al., 2020, 2021). SIDU creates a saliency map based on the similarities, differences, and uniqueness values from gradients of the last convolutional layer in the CNN (Muddamsetty et al., 2021). This overcomes difficulties with RISE choosing random perturbations rather than finding ones that provide the best explanation (Muddamsetty et al., 2021).

5.7.2. Method analysis

A major advantage of occlusion tests is that they are a post-hoc black box explainability method, which means that they are not limited to

Table 7
Domain summary for LIME.

Title and reference	Domain	Application
"Why Should I Trust You?": Explaining the Predictions of Any Classifier (Ribeiro et al., 2016)	Explainability of CNN decisions	Using explainability methods to increase user trust in machine learning models
Local Interpretable Model-agnostic Explanations for Music Content Analysis (Mishra et al., 2017)	Music	Classification of audio snippets
DLIME: A Deterministic Local Interpretable Model-agnostic Explanations Approach for Computer-aided Diagnosis Systems (Zafar and Khan, 2019)	Medical diagnoses	Using agglomerative Hierarchical Clustering to train the model, rather than random perturbation to increase model trust
Artificial Neural Networks Interpretation using LIME for Breast Cancer Diagnosis (Hakkoum et al., 2020)	Medical diagnoses	Applying LIME to the MLP and Deep MLP neural networks
OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms (Visani et al., 2020)	Medical diagnoses	Improving the stability of LIME explanations
Detection of COVID-19 Findings by the Local Interpretable Model-agnostic Explanations Method of Types-based Activations Extracted from CNNs (Togacar et al., 2022)	Medical diagnoses	Using LIME to explain COVID-19 diagnoses over other diseases with similar symptoms
Local Interpretable Model-agnostic Explanations for Classification of Lymph Node Metastases (Palatnik de Sousa et al., 2019)	Medical diagnoses	Using LIME to explain CNN decisions on lymph node metastases
Evolving Local Interpretable Model-agnostic Explanations for Deep Neural Networks in Image Classification (Wang et al., 2021)	Image recognition	Combining LIME with genetic algorithms to improve its computational efficiency

Fig. 14. Occlusion of different parts of an image via alternate image overlap.
Source: Adapted from Osherov and Lindenbaum (2017).

one neural architecture (Petsiuk et al., 2018) . This portability between neural networks could increase the speed with which confidence is gained in machine learning models for safety–critical applications. Additionally, methods such as TimeREISE expand the type of data analyzed to time series data, which could increase the domain reach of these explanations.

However, as discussed in the section on LIME, black box explainability methods limit the true understanding of how a decision is made. Occlusion tests use random sub-sampling of images to produce an explanation, resulting it the explanation being non-deterministic (Visani et al., 2020).

5.7.3. Applications

Table 8 summarizes the sources used in this section on occlusion tests. The summary includes the domain of the source and the application of the model.

5.8. Shapley Additive Explanations (SHAP)

5.8.1. Method description

Shapley Additive Explanations (SHAP) is another black box explainability method that produces an explanation after the decision has been made (Zhao et al., 2020). SHAP looks at individual feature inputs and analyzes their importance in the final decision (Mangalathu et al., 2020; Rashed-Al-Mahfuz et al., 2021). This method was designed using co-operative game theory to linearly map the non-linear decision (Mangalathu et al., 2020; Young et al., 2019). This is done by sub-sampling the feature inputs in varying combinations to understand their interactions and effects on the decision (Young et al., 2019). The result of this method is a contribution value for each feature input towards the output decision (Mangalathu et al., 2020) (Young et al., 2019). The feature sub-sampling makes SHAP a good method for both local and global explanations based on which features are sampled (Mangalathu et al., 2020).

Table 8
Domain summary for occlusion tests.

Title and reference	Domain	Application
SIDU: Similarity Difference and Uniqueness Method for Explainable AI (Muddamsetty et al., 2020)	Explaining deep learning decisions	Applying explainable AI (XAI) to safety–critical domains
Increasing CNN Robustness to Occlusions by Reducing Filter Support (Osherov & Lindenbaum, 2017)	CNN training	Training a CNN on partially occluded data to increase robustness
Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease (Nigri et al., 2020)	Medical diagnoses	Using occlusion heatmaps to avoid false positive Alzheimer's Disease diagnoses
A Novel Explainable Machine Learning Approach for EEG-Based Brain-Computer Interface Systems (Ieracitano et al., 2021a,b)	Medical diagnoses	Analyze which cortical regions affect hand movement planning the most
Deep Learning on Fundus Images Detects Glaucoma Beyond the Optic Disc (Hemelings et al., 2021)	Medical diagnoses	Reduce false negative diagnoses without increasing false positive diagnoses
Toward an Augmented and Explainable Machine Learning Approach for Classification of Defective Nanomaterial Patches (Ieracitano et al., 2021a,b)	Nanomaterials	Identifying defective and non-defective nanofibers in nanomaterials
Improving Performance of the PRYSTINE Traffic Sign Classification by Using a Perturbation-Based Explainability Approach (Sudars et al., 2022)	Autonomous driving	Improving traffic sign recognition to overcome obstacles such as partial occlusion and weather conditions
RISE: Randomized Input Sampling for Explanation of Black-box Models (Petsiuk et al., 2018)	Explainability of image recognition decisions	Explaining CNN image recognition decisions
TimeREISE: Time Series Randomized Evolving Input Sample Explanation (Mercier et al., 2022)	Explaining CNN decisions	Explaining CNN decisions in safety–critical domains
Visual Explanation of Black-box Model: Similarity Difference and Uniqueness (SIDU) Method (Muddamsetty et al., 2021)	Image recognition	Improving the accuracy of the explanation over RISE

There are several variations of SHAP that approximate the values in a computationally cheaper way, such as Kernel SHAP and Deep SHAP (Mangalathu et al., 2020; Young et al., 2019). Kernel SHAP reduces the computational load by combining Linear LIME and Shapely values to linearize the explanation (Lundberg and Lee, 2017). SHAP can also be combined with other explainability methods to reduce the computational requirements. For example, combining SHAP with the DeepLIFT method, a method similar to LWRP, resulting in Deep SHAP (Lundberg and Lee, 2017). Deep SHAP computes the SHAP value for each input separately by assuming that the CNN model is linear (Lundberg and Lee, 2017). These two variations of SHAP, as well as others, allow SHAP to be used for explainability despite its large computational requirement.

5.8.2. Method analysis

One of the advantages of SHAP is that the feature sub-sampling allows it to accurately describe decisions both locally and globally. Additionally, it is a black box explainability method, which gives it the flexibility to work on multiple neural network architectures. However, a major disadvantage of using SHAP values for explainability is that calculating them takes significant processing power when given many input features (Young et al., 2019). Since the method sub-samples the feature inputs and tests combinations of them, the processing power increases exponentially with the number of feature inputs (Young et al., 2019). This disadvantage is, however, addressed in some of the SHAP variations.

5.8.3. Applications

Table 9 summarizes the sources used in this section on SHAP. The summary includes the domain of the source and the application of the model.

5.9. Novel methods

The methods discussed earlier are widespread, but this section discusses some less common methods for explainability of CNNs. These methods include explanation vectors, domain knowledge infusion, decision trees, and explanatory graphs.

5.9.1. method—explanation vectors

Local explanation vectors use a Bayes classifier and Gaussian process, which is the derivative of the image function (Baehrens et al., 2010; Guidotti et al., 2019). This produces a vector field, as shown in Fig. 15 where (a) is the training data, (b) is the positive class function for the model's decision, (c) shows the local explanation vectors supporting the decision, and (d) shows all of the explanation vectors, including those that do not support the decision, which point away from the object (Baehrens et al., 2010). The vectors are gradients that represent the probability of the region being part of their defined class of objects (Guidotti et al., 2019; Rasmussen et al., 2018). Vectors with larger magnitudes contribute more to the decision of the CNN than those with smaller magnitudes (Baehrens et al., 2010). Although this paper focuses on the use of explanation vectors in explaining the decisions of a CNN, explanation vectors can explain the decision of any classification method (Guidotti et al., 2019).

5.9.2. Method analysis—explanation vectors

Explanation vectors have the advantage that they work for any classification method, not just CNNs (Baehrens et al., 2010) . Therefore, building confidence in this explainability method could be portable to other machine learning algorithms and automatically increase confidence in those algorithms. However, as seen in Fig. 15, the explanation

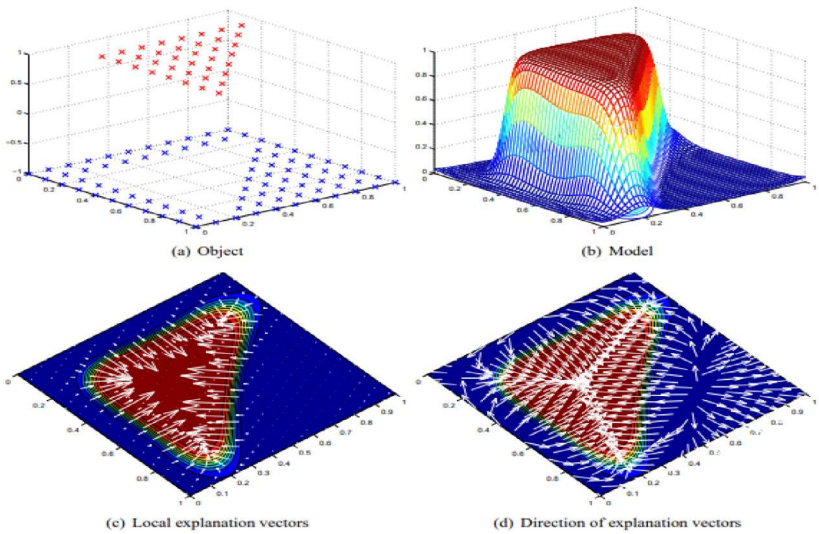


Fig. 15. Local explanation vectors.
Source: Adapted from Baehrens et al. (2010).

Table 9
Domain summary for SHAP.

Title and reference	Domain	Application
SHAP Values for Explaining CNN-based Text Classification (Zhao et al., 2020)	Text classification	Computing local and global SHAP values for CNN text classification models
Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive exPlanations (SHAP) Approach (Mangalathu et al., 2020)	Structural engineering	Seismic damage assessment of reinforced concrete columns and shear walls
Deep Neural Network or Dermatologist? (Young et al., 2019)	Medical diagnoses	Explainability of CNN decisions in melanoma diagnoses
A Deep Convolutional Neural Network Method to Detect Seizures and Characteristic Frequencies Using Epileptic Electroencephalogram (EEG) Data (Rashed-Al-Mahfuz et al., 2021)	Medical diagnoses	Improving confidence in CNN classifiers for diagnosis of epileptic seizures
A Unified Approach to Interpreting Model Predictions (Lundberg and Lee, 2017)	Explaining deep learning decisions	Combining explainability methods to reduce the computational power needed to compute Shapley values

Table 10
Domain summary for explanation vectors.

Title and reference	Domain	Application
How to Explain Individual Classification Decisions (Baehrens et al., 2010)	Explaining black box models	Useful for explaining any classification method
A Survey of Methods for Explaining Black Box Models (Guidotti et al., 2019)	Explainability of black box models	Correlating explainability methods with the black box model that they were designed to explain
Visualization of Nonlinear Classification Models in Neuroimaging (Rasmussen et al., 2018)	Neuroimaging	Analyzing functional MRIs (fMRIs) for area of focus when shown different visual stimuli

is mathematical and could be difficult for a non-mathematical expert to follow, which decreases the number of domains that it could be useful in.

5.9.3. applications—explanation vectors

Table 10 summarizes the sources used in this section on explanation vectors. The summary includes the domain of the source and the application of the model.

5.9.4. method—domain knowledge infusion

This method increases confidence in a CNN's decision by comparing the original decision with one based on a simpler data set created with domain knowledge, which is a dataset most relevant for classification (Islam et al., 2019, 2020a,b). Using domain knowledge can simplify complex regions of the image which reduces the number of features that can be used to classify an image, so a decision can be related back to a finite set of reasons (Islam et al., 2019). By using a finite set of known

Table 11
Domain summary for domain knowledge infusion.

Title and reference	Domain	Application
Infusing Domain Knowledge in AI-Based "Black Box" Models for Better Explainability with Application in Bankruptcy Prediction (Islam et al., 2019)	Banking	Bankruptcy prediction
Domain Knowledge Aided Explainable Artificial Intelligence for Intrusion Detection and Response (Islam et al., 2020a,b)	Cybersecurity	Using artificial intelligence to predict and respond to cybersecurity threats
Implications of Combining Domain Knowledge in Explainable Artificial Intelligence (Islam and Eberle, 2021a)	Cybersecurity	Using domain knowledge to explain artificial intelligence decisions in cybersecurity

features, there is higher confidence in the decision using this method, if the key features were correctly simplified (Islam et al., 2019). There are two parts to this classifier. First, the generalizer uses pattern mining to create a small set of standard features from a carefully selected and relevant training set that can be used to classify an image (Islam et al., 2019, 2020a,b; Islam and Eberle, 2021a). Second, the evaluator takes the original decision made by the CNN and compares it to the decision made by the CNN using the domain knowledge feature set (Islam et al., 2019, 2020a,b; Islam and Eberle, 2021a). If they match, then not only does the feature set explain the decision made by the CNN, but it also builds confidence in the decision by matching the two decisions (Islam et al., 2019). A mismatch between the decisions would indicate a low confidence in the decision made.

5.9.5. Method analysis—domain knowledge infusion

Domain knowledge infusion reduces the number of features that contribute to the decision, which makes the explanation easier to understand because there is a finite set of answers. Confidence in the model will likely increase if the explanation is easy to understand. Additionally, this method involves making two classification decisions and comparing the answers, which should increase confidence in the decision because two separately trained neural networks made the same decision.

However, this could cause issues if the features are over-simplified. For example, the CNN could recognize an ear, but that would be an overly simplified explanation of why the CNN classified the image as a dog over a cat. Experiments have shown that using all features for classification has slightly better accuracy than using select features (Islam et al., 2020a,b; Islam and Eberle, 2021a). Additionally, this method could run the risk of incorrectly simplifying the regions during classification, leading to a completely different decision than the original CNN.

5.9.6. applications—domain knowledge infusion

Table 11 summarizes the sources used in this section on domain knowledge infusion. The summary includes the domain of the source and the application of the model.

5.9.7. method—decision trees

Tree data structures format decision trees so that each path down the tree leads to a certain decision and tracing back up through the tree explains that decision (Zhang et al., 2019). A decision tree explains the CNN's decision in a coarse-to-fine direction (Zhang and Zhu, 2018c; Zhang et al., 2019). Filters in the CNN comprise the decision tree and sort through object characteristics in the image. For example, different appearances and positions of a bird's head and wings may help determine if the bird is standing or flying (Zhang et al., 2019). Higher-level nodes of the decision tree combine common characteristics of an object that are used in broadly identifying an image, such as

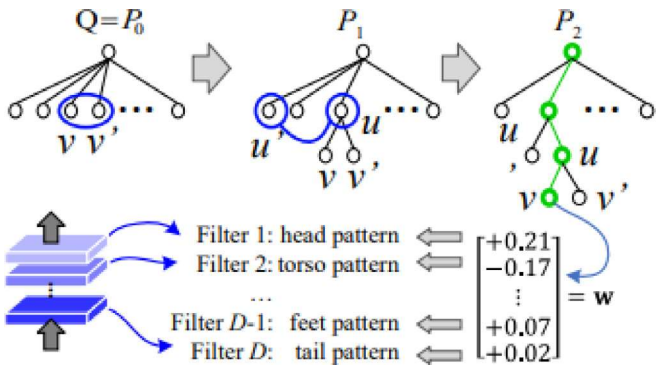


Fig. 16. Decision tree explanation.
Source: Adapted from Zhang et al. (2019).

Table 12
Domain summary for decision trees.

Title and reference	Domain	Application
Visual Interpretability for Deep Learning: A Survey (Zhang and Zhu, 2018c)	Image recognition	A survey of explainability methods for CNNs
Interpreting CNNs via Decision Trees (Zhang et al., 2019)	Image recognition	Explaining CNN image recognition decisions

recognizing a bird. Leaf nodes use finer details, such as feather coloring on the head, to decide the specific bird species (Zhang et al., 2019). Fig. 16 shows a visualization of building a decision tree where higher-level nodes combine common decision points and each path down the tree leads to a certain decision (Zhang et al., 2019).

The probabilities seen in the lower right corner of Fig. 16 represent the contribution of each node towards the decision (Zhang et al., 2019). These probabilities are found using analysis of filter activation from the CNN and help to explain the decision (Zhang et al., 2019). Note that the higher-level nodes have much heavier contributions, both for and against, than the finer details of the image. The training process for decision trees is simpler than some other training processes because the computer groups similar characteristics regardless of the presence of annotations made during training. However, as a result, the explanation is not as detailed as some other explainability methods (Zhang et al., 2019).

5.9.8. Method analysis—decision trees

Decision trees map the path of the CNN's decision. They show the breakdown of the decision through the CNN layers to clarify the decision process, which could increase confidence in the model. Feature grouping simplifies model training, but also reduces the explanation detail compared to other explainability methods. Additionally, the decision tree method assumes that all the possibilities, including those not chosen, are depicted in the tree. Since each node depicts a decision, it should show all possible choices (Zhang et al., 2019). Confidence could decrease proportionally to every probable solution not depicted in the decision tree.

5.9.9. applications—decision trees

Table 12 summarizes the sources used in this section on decision trees. The summary includes the domain of the source and the application of the model.

5.9.10. method—explanatory graphs

The use of graphs can lead to an interpretation of what each layer of the CNN filters. The graphs, made using a graph data structure, have layers corresponding to the layers of the CNN and the nodes in the

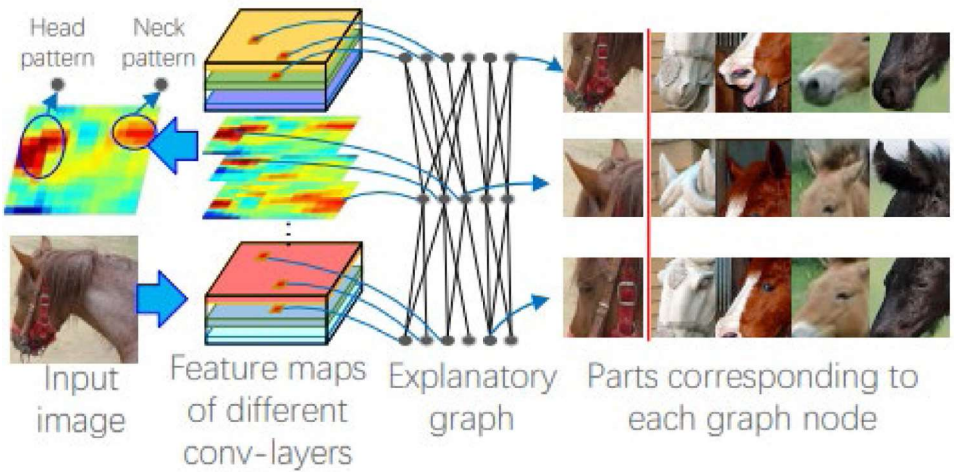


Fig. 17. Explanatory graph.
Source: Adapted from Zhang et al. (2018b).

graph represent the features discovered in each layer (Zhang et al., 2018b). While layers of the CNN may filter multiple features at once, the explanatory graph separates them to make the decision easier to understand (Zhang and Zhu, 2018c; Zhang et al., 2018b). Especially as the CNN reaches higher levels during the decision, multiple features in the image may activate neurons so the graph method helps to separate those features with a node representing each of the features, as shown in Fig. 17 (Zhang and Zhu, 2018c; Zhang et al., 2018b). The bottom of the figure shows the input image. As they go through the early convolutional layers, the characteristics of the image convert to nodes in the explanatory graph, which trace to their respective regions of the image. Going up in the figure, the input to a later convolutional layer is the feature map from the previous layer. As a feature map goes through a convolutional layer, the CNN will still filter out characteristics, which form nodes in the graph and correspond to regions in the image.

5.9.11. Method analysis—explanatory graphs

Explanatory graphs separate each layer and represent each node as an important feature. This makes the explanation detailed and the decision process easy to follow. Confidence could be higher using this explainability method because it is possible to follow each step in the decision process (Zhang et al., 2018b) . However, each node represents a feature in the image that contributes to the decision made. Depending on how detailed and specific the feature is, as well as if the neural network is in one of the later layers, it is possible that the features could be difficult to distinguish to the human eye, since the analyzed patterns are 70 × 70 pixels (Zhang et al., 2018b).

5.9.12. applications—explanatory graphs

Table 13 summarizes the sources used in this section on explanatory graphs. The summary includes the domain of the source and the application of the model.

Table 13
Domain summary for explanatory graphs.

Title and reference	Domain	Application
Visual Interpretability for Deep Learning: A Survey (Zhang and Zhu, 2018c)	Image recognition	A survey of explainability methods for CNNs
Interpreting CNN Knowledge Via an Explanatory Graph (Zhang et al., 2018b)	Image recognition	Explainability of CNNs

6. Source metadata

In total, this literature review used 90 sources. Analyzing the metadata of the sources showed some interesting trends. This section introduces some graphs and analysis of the sources regarding the publishing year, domain, publisher, and explainability method. Fig. 18 is a graph showing the frequency of publications by year. The data indicates that machine learning explainability research has gained popularity in more recent years given the effort to incorporate explanations into machine learning's classifications.

Fig. 18 can be broken down further to show the number of papers written for each explainability method by year, as shown in Fig. 19. The explainability methods available as early as the 1990s were logic rules and inversion. Deconvolution was a popular explainability method in the 2010s, and current popular methods are domain knowledge infusion, sensitivity analysis, and heatmaps, although other methods are still relevant and utilized.

Fig. 20 breaks down the percentages of each domain covered in the 85 sources. General explainability, healthcare and image recognition are the largest categories, but with versatile and flexible methods of explainability, CNNs could apply to more specific domains. The figure summarizes the other domains.

Fig. 21 breaks down the sources into percentages by publisher or conference. The two largest percentages are IEEE and arXiv. ACM also has several published papers on this topic with the rest of the publishers and conferences being roughly equal.

Fig. 22 shows the explainability methods by percentage. The most predominent methods are deconvolution, logic rules, LIME, and occlusion tests; however, there is growing interest and research in novel or underused methods. Some sources used in this literature review focused on using CNNs and did not use an explainability method. These sources are not included in Fig. 22.

7. Future work

This literature review discusses some of the most popular explainability methods that are currently available for CNN decisions, but this leaves room for future research. This section gives examples of potential directions for future research.

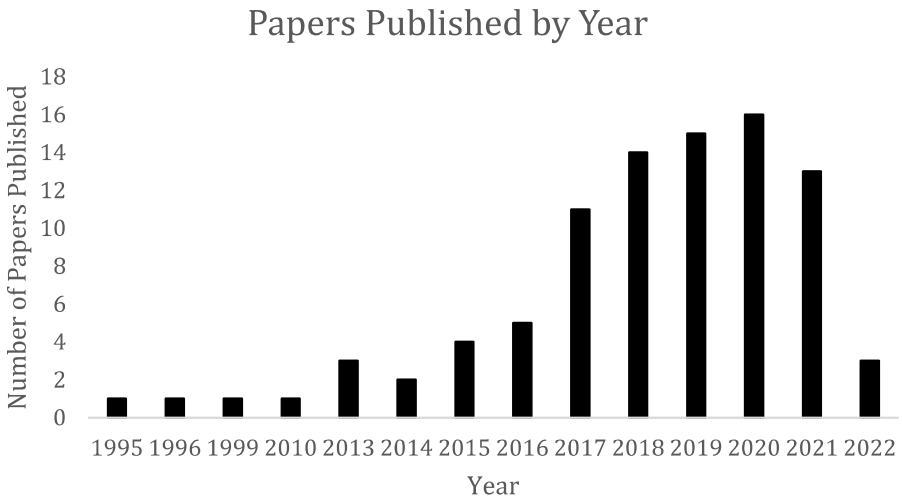


Fig. 18. Frequency of papers published by year.

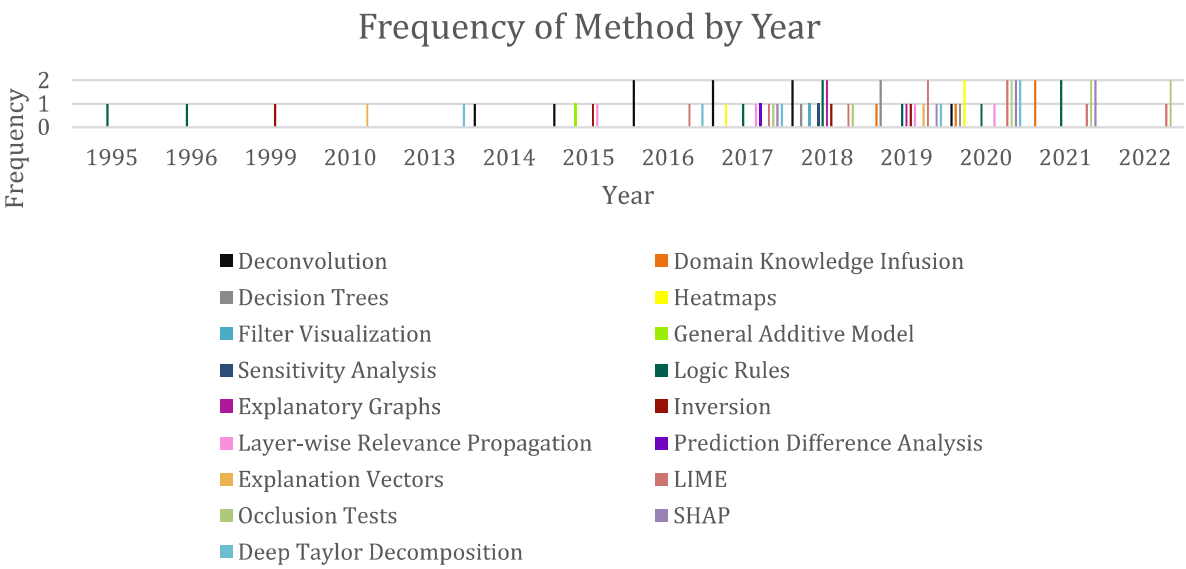


Fig. 19. Frequency of papers written for each method by year.

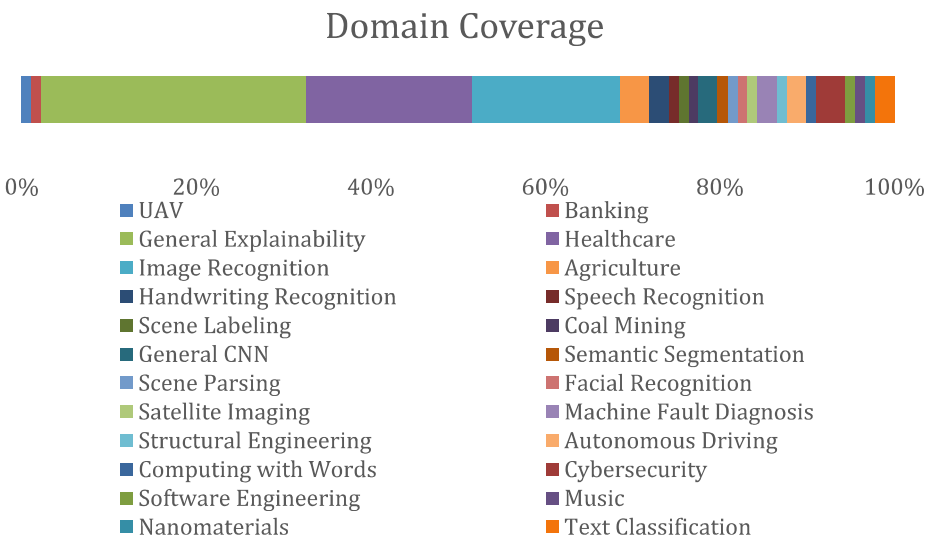


Fig. 20. Domain coverage by percentage of total sources.

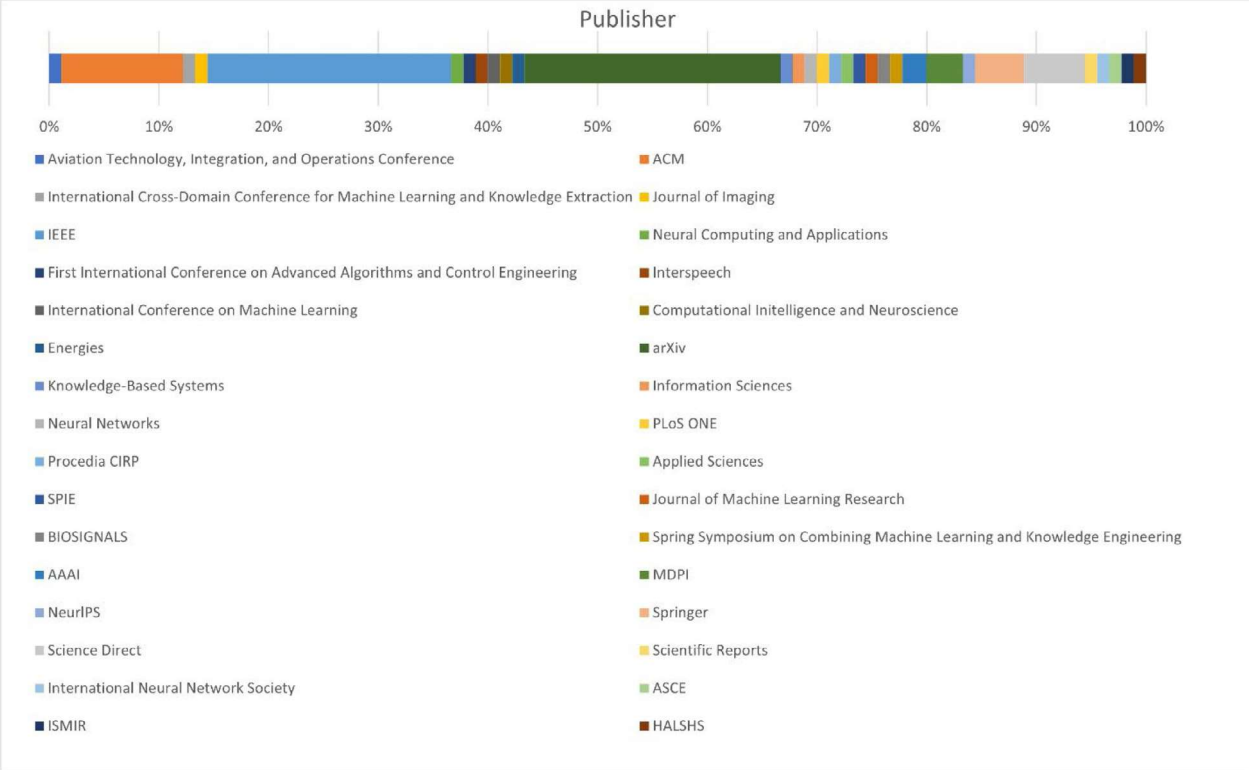


Fig. 21. Publishers by percentage.

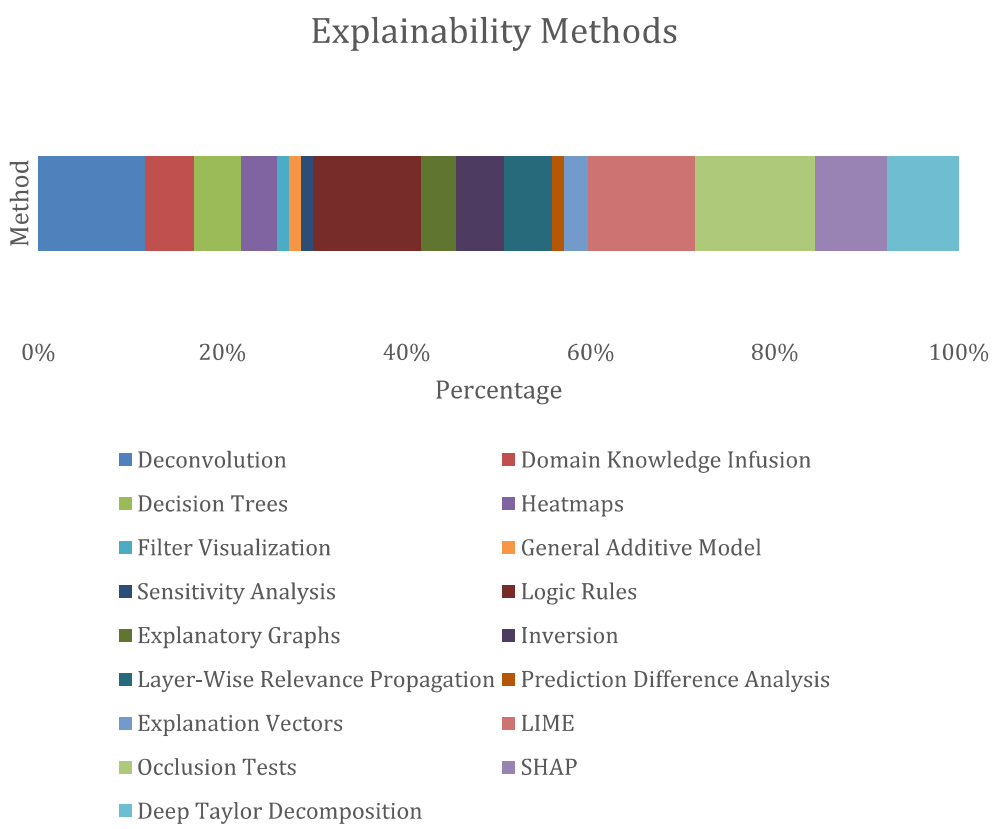


Fig. 22. Explainability methods by percentage.

7.1. Domain expansion and method mapping

One possible direction for future research is to explore additional domains in which CNNs are used and specifically the usage of explainability methods within those domains. The goal of this research would be to map which explainability methods are most effective for each domain found. As discussed earlier, some methods may be preferable to others because of the necessary level of mathematical expertise. Additionally, an explainability method, such as logic rules, may not provide the best explanation in a domain where a visual explanation is desired, such as in tumor recognition. Gathering additional data on domains and corresponding methods could begin to map domains to their most effective methods.

7.2. Ontology development

Using the mapping of domains and their most effective methods enables the development of an ontology of explainability methods for CNNs. An ontology is a mapping of concepts in a domain and their relations among each other (Valja et al., 2020). The goal of this ontology is that experts using a CNN in any listed domain can reference the ontology to find the most effective explainability method in their domain. This ontology could also suggest combinations of methods if there are gaps in the explanation that the domain would benefit from filling. Experts using a CNN can view the advantages and disadvantages of each of the available methods and choose the one that fits their application the best.

7.3. Method expansion

This literature review does not encompass all available CNN explainability methods. Future work could include expanding the methods researched and mapping them to domains that they are commonly used in. This research could also be added to the domain and method ontology. Finding more perturbation-based and approximation-based methods could make this paper a more complete review of the available methods, their advantages, and their disadvantages.

7.4. Method development

The purpose of these explainability methods is to increase confidence in the accuracy of the CNN decisions so that industries can use them, specifically in high-risk fields such as healthcare, banking, and autonomous driving. Reaching this goal is elusive, but the next steps towards increasing confidence are to improve the already effective explainability methods. A possible improvement is to combine some of the unique advantages of each method, discussed in the analysis sections of the survey, to improve confidence. For example, deconvolutional neural networks can densify the image after each unpooling layer, LWRP combines both positive and negative evidence to make a more complete explanation of the decision, and explanatory graphs separate characteristics into their own node on the graph to make it easier to trace the explanation back to the input image. Combining certain advantages could be helpful to specific domains, depending on the type of explanation that domain would benefit from. Technical research into the feasibility of combining these features into a single explanation model could be a next step towards building confidence in CNNs.

7.5. Subjectivity in decision making

While neural networks attempt to imitate the human decision process, code cannot imitate certain aspects of human decisions, which refers to incompleteness (Burkart and Huber, 2021) . Humans have emotions that play a role in their decision-making process that make it difficult to ensure a human-like decision. Some developed models predict human emotions to gain a better understanding of how they would react, but this area requires more research (Usman et al., 2021). Future work in this area could be further development of the human emotion models. The challenge in this lies in the testing and model verification. The standard testing section discusses this in more detail.

7.6. Standard testing

In addition to development of a new explainability method, a standard for testing the explainability methods needs to be established (Burkart and Huber, 2021). A method might explain the decision made by the CNN, but a standard test for accuracy for that explainability method needs to be developed. Biases in human opinions, even those of experts, affect the accuracy of an explainability method test, so a standard test to determine the accuracy of each method in relation to others could improve the confidence of both the CNN decision and the explainability method. Such a testing technique may not be identical for all domains or methods. For example, analyzing the accuracy of a heatmap may be different from analysis of explanation vectors because of the format difference. There is a need to develop a standard accuracy test for each method. The rigor of the testing technique could depend on the safety-criticality of the domain.

8. Conclusion

This paper is a literature review and analysis of explainability methods for CNNs. The paper uses the findings of 90 papers to summarize several popular and novel explainability methods for CNN decisions. The survey includes 12 explainability methods along with several of their benefits and disadvantages. Although CNNs can be accurate, they do not explain the reasons for their decisions, so confidence in their accuracy is lacking too much to use them in high-risk applications, such as healthcare and banking. This paper discusses deconvolutional neural networks, LWRP, deep Taylor decomposition, logic rules, inversion, LIME, occlusion tests, and SHAP which are common explainability methods. Additionally, the discussion includes explanation vectors, domain knowledge infusion, decision trees, and explanatory graphs, which are less common novel methods. Each of these methods have advantages that some of the others do not. The analysis of benefits and disadvantages of each method, as well as the tables in each method section mapping that method to domains is a unique contribution, not found in related work. The table in each method section summarizes the sources found discussing that method. It includes the domain and specific application that each source used the explainability method in and the result of the study. Expanding the analysis and tables written in this paper can lead to a more in-depth ontology of explainability methods as a reference for domain experts using a CNN to find the most effective method for their domain. Future research could also include exploring the feasibility of combining some of these advantages into new explainability methods. Using such methods could help to bridge the gap towards using CNNs in high-risk applications.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Lynn Vonder Haar: Conceptualization, Investigation, Writing – original draft. **Timothy Elvira:** Validation, Writing – review & editing. **Omar Ochoa:** Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Omar Ochoa reports a relationship with Concepts Beyond that includes: funding grants.

Data availability

No data was used for the research described in the article.

References

Abdel-Hamid, O., et al., 2013. Exploring convolutional neural network structures and optimization techniques for speech recognition. Interspeech 11, 73–75.

Albawi, S., et al., 2017. Understanding of a convolutional neural network. In: International Conference on Engineering and Technology. Akdeniz University, Turkey.

Andrews, R., et al., 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl.-Based Syst. 8 (6), 373–389.

Bach, S., et al., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10 (7).

Baehrens, D., et al., 2010. How to explain individual classification decisions. J. Mach. Learn. Res. (11), 1803–1831.

Bazen, S., Joutard, X., 2013. The Taylor Decomposition: A Unified Generalization of the Oaxaca Method To Nonlinear Models. HALSHS.

Blanco-Justicia, A., Domingo-Ferrer, J., 2019. Machine learning explainability through comprehensivle decision trees. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Canterbury, UK.

Bologna, G., 2019. A simple convolutional neural network with rule extraction. Appl. Sci. 9 (12).

Bonanno, D., et al., 2017. An Approach to Explainable Deep Learning using Fuzzy Inference. SPIE Defense and Security, Anaheim, CA.

Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised machine learning. J. Artificial Intelligence Res. 70, 245–317.

Cao, F., et al., 2020. Deconvolutional neural network for image super-resolution. Neural Netw. 132, 394–404.

Caruana, R., et al., 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. Association for Computing Machinery, Sydney NSW Australia.

Dolph, C.V., et al., 2018. Towards explainability of UAV-based convolutional neural networks for object classification. In: Aviation Technology, Integration, and Operations Conference. Atlanta, GA.

Du, M., et al., 2018. Towards explanation of DNN-based prediction with guided feature inversion. arXiv preprint.

Fan, F.-L., et al., 2021. On interpretability of artificial neural networks: A survey. IEEE Trans. Radiat. Plasma Med. Sci. 741–760.

Flutura, S., et al., 2020. Interactive machine learning and explainability in mobile classification of forest-aesthetics. In: EAI International Conference on Smart Objects and Technologies for Social Good (90–95). Association for Computing Machinery, Aveiro.

Gao, H., et al., 2017. Pixel deconvolutional networks. arXiv.

Grezmak, J., et al., 2019. Explainable convolutional neural network for gearbox fault diagnosis. Proc. CIRP 80, 476–481.

Grezmak, J., et al., 2020. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis. IEEE Sens. J. 20 (6), 3172–3181.

Guidotti, R., et al., 2018. Local rule-based explanations of black box decision systems. arXiv preprint.

Guidotti, R., et al., 2019. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Hakkoum, H., et al., 2020. Artificial neural networks interpretation using LIME for breast cancer diagnosis. In: World Conference on Information Systems and Technologies (15–24). Springer.

Hemelings, R., et al., 2021. Deep learning on fundus images detects glaucoma beyond the optic disc. Sci. Rep..

Hoffman, R.R., et al., 2018. Metrics for explainable AI: Challenges and prospects. arXiv.

Ieracitano, C., et al., 2021a. A novel explainable machine learning approach for EEG-based brain-computer interface systems. Neural Comput. Appl..

Ieracitano, C., et al., 2021b. Toward an augmented and explainable machine learning approach for classification of defective nanomaterial patches. In: International Conference on Engineering Applications of Neural Networks (244–255). Springer, Crete.

Islam, S.R., Eberle, W., 2021a. Implications of combining domain knowledge in explainable artificial intelligence. In: Spring Symposium on Combining Machine Learning and Knowledge Engineering. Palo Alto, CA.

Islam, S., et al., 2019. Infusing Domain Knowledge in AI-Based Black Box Models for Better Explainability with Application in Bankruptcy Prediction. Association for Computing Machinery.

Islam, S.R., et al., 2020a. Domain knowledge aided explainable artificial intelligence for intrusion detection and response. arXiv preprint.

Islam, S.R., et al., 2020. Towards quantification of explainability in explainable artificial intelligence methods. In: The Thirty-third International Flairs Conference.

Islam, S., et al., 2021b. Explainable artificial intelligence approaches: A survey. arXiv preprint.

Jensen, C.A., et al., 1999. Inversion of Feedforward Neural Networks: Algorithms and Applications. IEEE.

Kauffmann, J., et al., 2020. Towards explaining anomalies: A deep taylor decomposition of one-class models. Pattern Recognit..

Kindermans, P.-J., et al., 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. In: NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems. Barcelona: arXiv.

Kitchenham, B., Brereton, P., 2013. A systematic review of systematic review process research in software engineering. Inf. Softw. Technol. 2049–2075.

Lin, Z., et al., 2019. Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explanability algorithms. arXiv.

Liu, Y., 2018b. Feature extraction and image recognition with convolutional neural networks. In: First International Conference on Advanced Algorithms and Control Engineering. National Pingtung University of Science and Technology, Taiwan.

Liu, H., et al., 2018a. Learning deconvolutional deep neural network for high resolution medical image reconstruction. Inform. Sci. 468, 142–154.

Liu, Y., et al., 2021. Synthetic benchmarks for scientific research in explanable machine learning. arXiv.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems.

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA.

Mangalathu, S., et al., 2020. Failure mode and effects analysis of RC members based on machine-learning-based shapley additive explanations (SHAP) approach. Eng. Scructures.

Mercier, D., et al., 2022. Timereise: Time series randomized evolving input sample explanation. Sensors.

Mishra, S., et al., 2017. Local interpretable model-agnostic explanations for music content analysis. ISMIR 537–543.

Mohan, R., 2014. Deep deconvolutional networks for scene parsing. arXiv preprint.

Montavon, G., et al., 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognit. 211–222.

Montavon, G., et al., 2019. Layer-wise relevance propagation: An overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. pp. 193–209.

Mousavi, N., et al., 2016. Understanding how deep neural networks learn face expressions. In: International Joint Conference on Neural Networks. Vancouver, Canada.

Muddamsetty, S.M., et al., 2020. SIDU: Similarity difference and uniqueness method for explainable AI. In: International Conference on Image Processing.

Muddamsetty, S.M., et al., 2021. Visual explanation of black-box model: Similarity difference and uniqueness (SIDU) method. Comput. Vis. Pattern Recognit..

Nigri, E., et al., 2020. Explainable deep cnns for MRI-based diagnosis of Alzheimer's disease. In: International Joint Conference on Neural Networks, pp. 1–8.

Noh, H., et al., 2015. Learning deconvolution network for semantic segmentation. In: IEEE International Conference on Computer Vision. Santiago, Chile.

Odena, A., et al., 2016. Deconvolution and checkerboard artifacts. Distill.

Osherov, E., Lindenbaum, M., 2017. Increasing CNN robustness to occlusions by reducing filter support. In: IEEE International Conference on Computer Vision, pp. 550–561.

Palatnik de Sousa, I., et al., 2019. Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors.

Petsiuk, V., et al., 2018. RISE: Randomized input sampling for explanation fo black-box models. arXiv.

Pinheiro, P.O., Collobert, R., 2014. Recurrent convolutional neural networks for scene labeling. In: International Conference on Machine Learning. Beijing, China.

Priyadharshini, R., et al., 2019. Maize leaf disease classification using deep convolutional neural networks. Neural Comput. Appl. (31), 8887–8895.

Pu, Y., et al., 2019. Image recognition of coal and coal gangue using a convolutional neural network and transfer learning. Energies 12 (9).

Rashed-Al-Mahfuz, M., et al., 2021. A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (EEG) data. IEEE J. Transl. Eng. Health Med..

Rasmussen, P.M., et al., 2018. Visualization of nonlinear classification models in neuroimaging. BIOSIGNALS.

Ribeiro, M.T., et al., 2016. Why should I trust you?: Explaining the predictions of any classifier. In: International Conference on Knowledge Discovery (1135–1144). Association for Computing Machinery, New York.

Rokade, P., Alluri, B., 2021. Towards quantification of explainability algorithms. In: The 5th International Conference on Advances in Artificial Intelligence (ICAAI) (31–37). ACM.

Samek, W., et al., 2017a. Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. 28 (11), 2660–2673.

Samek, W., et al., 2017b. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal (1), 39–48.

Singh, A., et al., 2020. Explainable deep learning models in medical image analysis. J. Imaging 6 (6).

Sladojevic, S., et al., 2016. Deep neural networks based recognition of plant diseases by leaf image classification. Comput. Initell. Neurosci..

Sudars, K., et al., 2022. Improving performance of the PRYSTINE traffic sign classification by using a perturbation-based explainability approach. In: Formal Verification of Imaging Algorithms for Autonomous System.

Sun, X., et al., 2018. Image recognition of tea leaf diseases based on convolutional neural network. In: International Conference on Security, Pattern Analysis, and Cybernetics. Piscataway, NJ.

Tao, Y., et al., 2017. Unsupervised-restricted deconvolutional neural network for very high resolution remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55 (12), 6805–6823.

Togacar, M., et al., 2022. Detection of COVID-19 findings by the local interpretable model-agnostic explanations method of types-based activations extracted from CNNs. Biomed. Signal Process. Control.

Usman, M., et al., 2021. A human-in-the-loop probabilistic CNN-fuzzy logic framework for accident prediction in vehicular networks. IEEE Sens. J. 21 (14), 15496–15503.

Valja, M., et al., 2020. Automating threat modeling using an ontology framework. Cybersecurity.

Visani, G., et al., 2020. Optilime: Optimized LIME explanations for diagnostic computer algorithms. arXiv.

Wang, B., et al., 2021. Evolving local interpretable model-agnostic explanations for deep neural networks in image classification. In: Genetic and Evolutionary Computation Conference Companion (173–174). ACM.

Xi, Z., Panoutsos, G., 2018. Interpretable machine learning: convolutional neural networks with RBF fuzzy logic classification rules. In: International Conference on Intelligent Systems. Phuket, Thailand.

Yang, Z., et al., 2019. Neural network inversion in adversarial setting via background knowledge alignment. In: ACM SIGSAC Conference on Computer and Communications Security. London, UK.

Young, K., et al., 2019. Deep neural network or dermatologist?. arXiv.

Yuan, Y., 2017. Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation. arXiv preprint.

Zadeh, L.A., 1996. Fuzzy logic=computing with words. IEEE Trans. Fuzzy Syst. 4 (2), 103–111.

Zafar, M.R., Khan, N.M., 2019. DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Zurich, Switzerland.

Zhang, Q., Zhu, S.-C., 2018c. Visual interpretability for deep learning: a survey. arXiv preprint.

Zhang, Q., et al., 2018a. Interpretable convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT.

Zhang, Q., et al., 2018b. Interpreting CNN knowledge via an explanatory graph. In: AAAI Conference on Artificial Intelligence. New Orleans, LA.

Zhang, Q., et al., 2019. Interpreting CNNs via decision trees. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA.

Zhao, W., et al., 2020. SHAP values for explaining CNN-based text classification models. arXiv.

Zintgraf, L.M., et al., 2017. Visualizing deep neural network decisions: prediction difference analysis. In: International Conference on Learning Represenations. Toulon, France.