## Visual Explanation

### Backpropagation-based approaches

1. Simonyan et al. (2013) proposed visualization of partial derivatives of the output on pixel level.
2. Zeiler and Fergus (2014) introduced deconvolution for creating saliency maps.
3. Springenberg et al. (2014) developed guided backpropagation for saliency maps.
4. Zhou et al. (2016) introduced Class Activation Mapping (CAM) for local, model-specific, post hoc explanation.
5. Jiang et al. (2019) used an ensemble of CNNs and provided a weighted combination of CAMs for localization of diabetic retinopathy.
6. Lee et al. (2019b) constructed CAMs of the output of an ensemble of four CNNs for the detection of acute intracranial hemorrhage.
7. Liao et al. (2019) proposed multi-scale CAMs for better identification of small structures on fundus images of the retina.
8. Shinde et al. (2019a) developed ‘High Resolution’ CAMs for accurate localization of brain tumors on MRI.
9. García-Peraza-Herrera et al. (2020) proposed extracting CAMs at multiple resolutions for highlighting interpapillary capillary loop patterns in endoscopy images.
10. Selvaraju et al. (2017) introduced Gradient-weighted Class Activation Mapping (Grad-CAM) as a generalization of CAM.
11. Bach et al. (2015) introduced layer-wise relevance propagation (LRP) for assigning relevance scores to input neurons.
12. Lundberg and Lee (2017) proposed Deep SHapley Additive exPlanations (Deep SHAP) for explaining predictions using Shapley values.
13. Jetley et al. (2018) proposed a trainable attention mechanism for highlighting where the network pays attention to input images.
14. Schlemper et al. (2019) used trainable attention and introduced grid attention for capturing anatomical information in medical images.

### Perturbation-based approaches

1. Zeiler and Fergus (2014) used occlusion sensitivity analysis to visualize important parts of the image for classification.
2. Ribeiro et al. (2016) introduced Local Interpretable Model-agnostic Explanations (LIME) for providing local explanation by replacing complex models locally with simpler models.
3. Fong and Vedaldi (2017) introduced meaningful perturbation for detecting changes in predictions of a trained neural network.
4. Uzunova et al. (2019) proposed replacing pathological regions with healthy tissue equivalent using a variational autoencoder (VAE) for more meaningful perturbations.
5. Lenis et al. (2020) used inpainting to replace pathological regions with healthy tissue equivalents for better localization of pathology.
6. Zintgraf et al. (2017) adapted prediction difference analysis for generating saliency maps by assigning relevance values to each pixel.
7. Schwab et al. (2020) used a patch-based approach with multiple instance learning to localize critical findings in chest X-ray.
8. Araújo et al. (2020) used multiple instance learning to explain important areas of a fundus photograph for diabetic retinopathy assessment.

## Textual Explanation

### Image Captioning

1. Vinyals et al. (2015) proposed an end-to-end image captioning framework that combined a convolutional neural network (CNN) for image encoding and a long-short term memory (LSTM) network for textual encoding. They used human-generated sentences as ground truth for training and the bilingual evaluation understudy (BLEU) metric for evaluation.
2. Singh et al. (2019) used an image captioning framework to provide textual explanations for chest X-rays. They trained the LSTM using Global Vectors (GloVe) and the radiology variant RadGloVe, and evaluated their model using BLEU, METEOR, CIDER, and ROUGE metrics. They found that using both RadGloVe and GloVe led to higher performance in the generated radiology reports.

### Image Captioning with Visual Explanation

1. Zhang et al. (2017a) introduced a framework that combined image captioning with visual explanation using dual attention for both text and imaging. This approach facilitated high-level interactions between image and text predictions, resulting in visual attention maps corresponding with textual explanations in histology images.
2. Wang et al. (2018) used a similar approach for chest X-rays, showing that different parts of the textual explanation led to different areas of saliency mapping in the image. They demonstrated a saliency map of the chest with multiple regions corresponding to different radiological findings.
3. Lee et al. (2019a) presented image captioning with visual explanation for breast mammograms. They added a visual word constraint loss to the text-generating LSTM to ensure that the provided explanations followed the correct jargon of breast mammography reports. They found that adding this loss improved the textual explanation quality and linked the radiology reports to visual saliency maps.

### Testing with Concept Activation Vectors (TCAV)

1. Kim et al. (2018) introduced TCAV, which provided human-friendly linear explanations of the internal state of neural networks in terms of human-understandable concepts. The TCAV algorithm used user-defined sets of examples of a concept and random non-concept examples. They demonstrated the feasibility of TCAV on a medical image processing example by relating physician annotations such as ‘microaneurysm’ to diabetic retinopathy in fundus imaging.
2. Clough et al. (2019) used TCAV to identify cardiac disease in cine-MRI by classifying the latent space of a Variational Autoencoder (VAE). They showed which clinically known biomarkers were related to cardiac disease and reconstructed images with low peak ejection rate by adding the Concept Activation Vectors (CAVs) to the latent space.
3. Graziani et al. (2020) expanded on TCAV by introducing regression concept vectors, which indicated continuous-valued measures of a concept. They demonstrated that using regression concept vectors could explain why the network classified one area of a breast histopathology image as cancer and another as healthy based on the concepts ‘contrast’ and ‘nuclei area’.

### Other Textual Explanation Techniques

1. Shen et al. (2019) used a hierarchical semantic CNN to predict malignancy of lung nodules on CT. They classified five textual descriptions of image characteristics representative of lung nodule malignancy typically assessed by a radiologist. Although their hierarchical semantic CNN did not significantly outperform a normal CNN in predicting nodule malignancy, the method provided human-interpretable characteristics of the nodules.

## Example-based explanation

### Triplet Network

1. Hoffer and Ailon (2015) introduced the concept of a triplet network, consisting of three identical networks with shared parameters.
2. Peng et al. (2019) applied a triplet network for example-based explanation in colorectal cancer histology.
3. Yan et al. (2018) utilized a triplet network to extract 32,000 clinically relevant lesions from the entire body.
4. Codella et al. (2018) combined a triplet loss with global average pooling to provide example-based explanation and visual explanations in dermatology images of melanoma.

### Influence Functions

1. Wei Koh and Liang (2017) proposed using influence functions to explain which inputs from a training set a decision was based on.
2. Wang et al. (2019) applied influence functions to explain classifications of liver lesions on multiphase MRI and their association with radiological characteristics.

### Prototypes

1. Chen et al. (2019) proposed using typical examples as explanation (i.e., prototypes) in a method called “this-looks-like-that.”
2. Uehara et al. (2019) used prototypes to explain why a neural network classified patches of histology images as cancer or not-cancer.