EasyChair Preprint

№ 11570

Investigating Explainable AI for Enhanced Atherosclerosis Detection and Decision Transparency

Amel Laidi and Mohammed Ammar

December 19, 2023

Investigating Explainable AI for Enhanced Atherosclerosis Detection and Decision Transparency

Amel Laidi
Faculty of Engineering, LIMOSE Laboratory
M'Hamed Bougara University
Boumerdes, Algeria
a.laidi@univ-boumerdes.dz

Mohammed Ammar
Engineering Systems and Telecommunication Laboratory
University M'Hamed Bougara
Boumerdes, Algeria
m.ammar@univ-boumerdes.dz

***Abstract*— Atherosclerosis, a cardiovascular disease, is commonly diagnosed through non-invasive imaging methods like Coronary CT Angiography (CCTA). Deep learning algorithms have demonstrated remarkable potential in assisting with the classification of CCTA images. However, the inherent opacity in the decision-making processes of black-box AI models presents a significant challenge in the medical field. Healthcare professionals require clear insights into the rationale behind AI system recommendations. In this research, we investigate the advantages of Explainable AI (XAI) algorithms in the context of a previously established deep learning model for atherosclerosis classification from CCTA images. Our study not only highlights the capability of XAI in elucidating the model's decision-making process but also showcases its potential in identifying misclassified cases, thereby contributing to the refinement and enhancement of the model's performance.**

Keywords— deep learning, atherosclerosis, coronary CT angiography, Explainable Artificial Intelligence, GradCam.

I. INTRODUCTION

The extensive use of AI technology in the medical field, particularly in atherosclerosis screening, has drastically transformed disease diagnostics. Through the use of machine learning and deep learning techniques, AI has enabled early, precise, and comprehensive diagnosis of coronary atherosclerotic heart disease (CAD), a common cardiovascular disease with high morbidity, disability, and societal burden [1]. However, the prevalent reliance on Deep Neural Networks (DNNs) in AI models poses a substantial challenge regarding their explainability and transparency. This opacity in DNN models restricts the comprehension of how they reach diagnostic conclusions, which proves to be a significant obstacle in safety-critical medical domains. To address this gap, Explainable AI (XAI) methodologies have emerged to elucidate the decision-making processes of these models; they can aid in adhering to regulatory prerequisites and ethical standards, thus elevating the overall efficiency and impartiality of the system. [2]

Moreover, prior research has highlighted the importance of explainable AI in diverse fields, exemplified by works like [3]. The research conducted by Suryani et al. [4] emphasizes the critical role of AI-based tools in identifying lung tumors from chest X-ray images. The proposed methodologies in their work, such as Seg-Grad-CAM (Semantic Segmentation via Gradient-Weighted Class Activation Mapping), align with the objectives of our investigation in providing precise and accurate localization of lesions or abnormalities within medical images.

In this context, our research aims to leverage the advancements in explainable AI methodologies for the task of atherosclerosis screening through CCTA images, ensuring not only the accuracy of disease diagnosis but also fostering trust and understanding in AI-assisted medical decision-making. The model under study is one that has been previously developed [5], fine-tuned, and specialized for the specific task of atherosclerosis screening.

II. BACKGROUND

A. What is AI ? and why deep learning?

Artificial intelligence is a wide term that refers to all types of computer systems trained to perform tasks that are normally associated with human intelligence or abilities, such as perception, reasoning, learning, and decision-making. AI systems are designed to analyze large amounts of data, recognize patterns, and make predictions or decisions based on that data.

There are different types of AI systems, including:

- **Rule-based systems:** AI systems that use pre-defined rules to make decisions or perform tasks. They are limited to the rules that have been programmed into them and cannot adapt to new situations.
- **Machine learning systems** : AI systems that can learn from data and improve their performance over time. There are different types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning.
- **Deep learning systems** : a subset of machine learning systems that use artificial neural networks to analyze large amounts of data and learn from it. Deep learning systems are particularly effective in tasks that involve image and speech recognition, natural language processing, and other complex tasks.

Deep learning involves the use of artificial neural networks that consist of multiple layers of interconnected nodes, each of which performs a specific function in the processing of data. The input data is fed into the network, and the network gradually learns to recognize patterns in the data by adjusting the weights of the connections between nodes.

Deep learning has shown tremendous promise in computer vision and healthcare applications. It can analyze and interpret complex images, such as X-rays or MRI scans, with greater accuracy than traditional image analysis techniques. This ability has enabled medical professionals to diagnose diseases and conditions at an early stage, leading to better treatment outcomes. Additionally, deep learning has been used in healthcare to improve patient outcomes through personalized treatment plans. By analyzing vast amounts of patient data, deep learning algorithms can identify patterns and correlations that would be impossible for human doctors to detect. This has the potential to lead to more accurate diagnoses, better treatment plans, and ultimately, improved patient outcomes. Overall, the combination of deep learning, computer vision, and healthcare has the potential to revolutionize the way we diagnose and treat diseases, ultimately leading to better healthcare for everyone. [6]

B. *Deep learning for atherosclerosis detection from Coronary CT Angiography*

Atherosclerosis is a common cardiovascular disease that is characterized by the accumulation of fatty deposits in the walls of arteries. Coronary CT Angiography (CCTA) is the key technique for atherosclerosis screening, it is a non-invasive imaging technique that can be used to visualize the coronary arteries and detect the presence of atherosclerotic plaques [7]. However, interpreting CCTA images can be challenging and time-consuming, particularly when there are multiple plaques or complex plaque morphology.

Deep learning algorithms have been developed to assist in the classification of CCTA images by automatically identifying and characterizing atherosclerotic plaques based on their morphology, composition, and location within the coronary arteries. These algorithms can analyze large amounts of data and learn to recognize patterns that are indicative of different types of plaques, such as calcified or non-calcified plaques.

In a recent study [5], we created a deep learning model and trained it on a publicly available dataset.

Dataset :

The dataset we used is an open-source collection of Coronary CT Angiography images for screening atherosclerosis. It consists of Mosaic Projection View (MPV) images of 18 views of straightened coronary arteries from 500 patients, created by combining unique ray-traced projections. The dataset was partitioned into 300 training images, 100 testing images, and 100 validation images. To balance the dataset, the training images were augmented 6-fold. The validation dataset contains one randomly selected artery per normal case and one diseased case. [8]

Model:

The model we used was a pretrained Residual network (ResNet) [9] with 101 layers.

ResNet was first presented at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2015, It innovated by incorporating additional branches within its architecture. Specifically, one branch, known as the identity block, forwards information unchanged, while the other, the convolutional block, processes data akin to a standard layer. The unaltered data, referred to as the "residual," is combined with the original signal traversing the network without modification. This architectural division ensures that one branch solely transmits gradients without modifying them. Deep residual networks are constructed by stacking these blocks, enabling robust learning through powerful gradient propagation.

The concept of residual blocks is visually demonstrated in the reference Fig. 1**Erreur ! Source du renvoi introuvable.**. These blocks' core lies in the "jump connection," which defines the essence of the residual blocks.

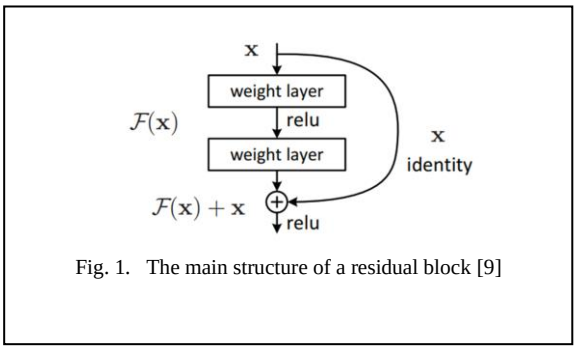

Fig. 1. The main structure of a residual block [9]

Residual Networks emerged as a solution to a prevalent issue encountered in deeper and more intricate networks—vanishing and exploding gradients. This phenomenon arises during the training of very deep networks when derivatives or slopes become exceedingly large or small, sometimes exponentially so, thereby complicating the training process.

There has been evidence that residual networks are easier to optimize and the gain in their accuracy is directly proportional with a considerable increase in their depth [10]. We used the resnet101 which has 101 layers.

Hyperparameters:

Mini-batch size : a segmented subset extracted from the training dataset via shuffling and partitioning. Its size can range from a single image to encompassing all examples within the training dataset.

Given the model's size, we trained on a mini-batch size of 64, which is neither too big nor too small. We used a learning rate of 10^{-4} which allowed the model to learn fast enough without missing local minima.

Dropout: a regularization technique employed during the training phase, where it randomly deactivates multiple

neurons within the network based on a pre-defined probability. This exclusion prevents their participation in both forward and backward propagation. Consequently, this alters the model's architecture during each iteration, fostering a more resilient training process. The implementation of Dropout notably enhances accuracy and substantially diminishes the required training time.

We used 50% dropout, which made the model more universal.

Number of training epochs: the number of cycles the model would go through the dataset. Training for too long epochs can cause the model to overfit, while not training enough can lead to underfitting.

The model trained for 20 epochs which made it perform well without overfitting.

The model took about 50 hours to train. It achieved 95.21% of accuracy, 90.48% positive predictive value, and 95.6% negative predictive value. The detailed results are illustrated in Fig. 2.

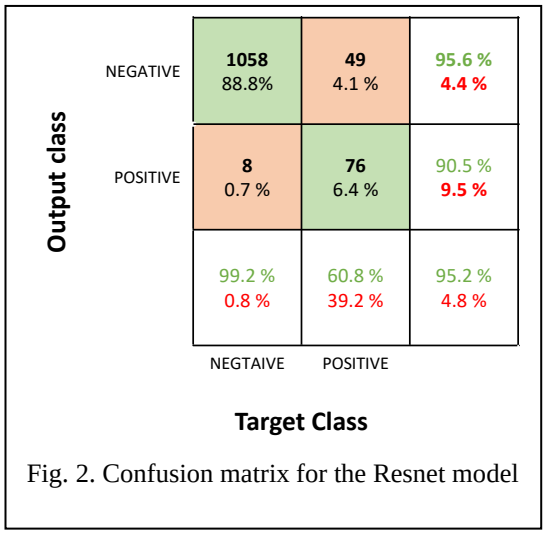

Fig. 2. Confusion matrix for the Resnet model

C. Challenges and limitations of black-box AI models

Black-box AI models, which refer to AI systems that are difficult or impossible to interpret, particularly Deep learning models, have become increasingly popular in the medical field for tasks such as atherosclerosis screening. However, there are several challenges and limitations associated with these models that need to be addressed to ensure their safe and effective use in clinical settings.

Not understanding the full process happening inside the model leads to a lack of transparency in the decision-making process. This can be particularly problematic in the medical field, if healthcare professionals cannot understand how an AI system arrived at a particular diagnosis or recommendation, they may be hesitant to act on it.

To address these challenges and limitations, recent research worked on developing AI models that are more transparent and interpretable, and to ensure that they are trained on unbiased and representative data. Additionally, there needs to be ongoing monitoring and evaluation of AI systems to ensure that they remain accurate and up to date with changes in the data and patient population. By addressing these challenges and limitations, the potential of black-box AI models to improve atherosclerosis screening and other medical tasks can be fully realized.

D. Explainable AI Algorithms

Explainable AI (XAI) algorithms are a type of AI system that is designed to be transparent and interpretable, allowing humans to understand how the AI system arrived at its decision or recommendation. XAI algorithms are intended to address the limitations of black-box AI models, which are often difficult or impossible to interpret.

There are several popular explainable AI algorithms that are used to provide transparency and interpretability to AI models. In this paper we took interest in three algorithms:

LIME (Local Interpretable Model-agnostic Explanations): LIME is a model-agnostic method for explaining the predictions of any black-box model. LIME works by creating a simpler, interpretable model that is locally faithful to the black-box model around the instance being explained. The simpler model can then be used to provide an explanation for the black-box model's prediction. [11]

LIME works by generating perturbations around the instance being explained and measuring the impact of these perturbations on the model's output. The algorithm then uses these perturbations to create a simpler, interpretable model that is locally faithful to the black-box model around the instance being explained.

LIME can provide a way to explain the prediction of any black-box model, without requiring knowledge of its internal workings. LIME is also a model-agnostic algorithm, meaning that it can be used with any type of model, including neural networks, decision trees, and support vector machines.

Grad-CAM (Gradient-weighted Class Activation Mapping): Grad-CAM is an algorithm for generating heatmaps that visualize the importance of different regions of an image for a given classification decision. Grad-CAM works by computing the gradients of the output class score with respect to the feature maps in the final convolutional layer of a neural network. The resulting gradient-weighted maps are then used to generate a final heatmap that visualizes the importance of different regions of an image for a given classification decision. [12]

The advantage of Grad-CAM over black-box models is that it provides a way to visualize the decision-making process of the neural network, which can be difficult to interpret using traditional methods. Grad-CAM also provides a way to identify which regions of an image are most important for the model's decision, which can be particularly useful in medical imaging.

Occlusion Sensitivity: Occlusion sensitivity is a technique for visualizing the importance of different regions of an image for a given classification decision by systematically occluding different parts of the image and

measuring the resulting change in the model's output. By comparing the model's output for the original image and the occluded images, it is possible to identify the most important regions of the image for the model's decision. [13]

Occlusion sensitivity is also a simple and easy-to-implement algorithm, making it a useful tool for interpreting the decisions of machine learning models.

E. XAI algorithms in atherosclerosis classification.

The use of explainable AI algorithms in atherosclerosis classification offers the typical advantages such as increased transparency and improved decision making, but it also adds another layer to the task of the classification. The main advantage of classifying the images directly, without first segmenting them, is the rapidity of the task. When using XAI algorithms, we can add a layer on the explanation, where the highlighted areas are most likely areas of high activity, in our case it could be the location of buildup. This not only allows for a quick detection of the disease, but it also guides the doctor to which region to investigate for higher risk of plaque.

III. RESULTS AND DISCUSSION

For a full analysis of the model, we ran the three algorithms on a number of images from the validation dataset. We tested true positives , true negatives, and false negatives.

False negatives were not part of the scope if this study, because they do not pose an urgent risk. A healthy patient that has been falsely classified as sick would be kept under medical care and eventually identified as healthy with further tests. While sick patients that were classified as healthy can be discharged without medical care, leading to further complications of their health.

A. True positive

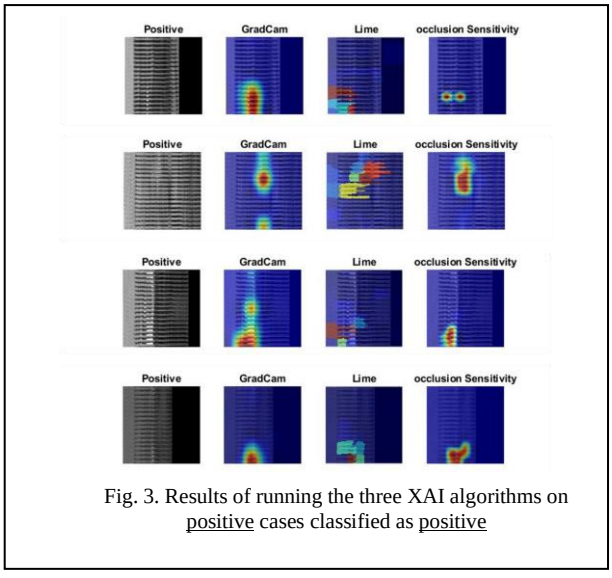

Fig. 3. Results of running the three XAI algorithms on positive cases classified as positive

True positive (TP) is a term used to describe a correctly predicted positive case. In other words, it is the case where the model correctly identifies a positive example as positive. In our case, the model has 90.48% positive predictive value, so it should be good at classifying positive cases.

At first sight, all three algorithms show similar activity at the same regions of the image (Fig. 3). This means that the deep learning model we used relied heavily on this region for its final decision.

The focus is on one small region, being highlighted in red, with no significant activity anywhere else.
The explainable models offer an extra layer of performance, it shows the region with higher probability of plaque and residue, which is useful for the rest of the medical care process.

B. True Negative (TN)

True negative (TN) is a term used to describe a correctly predicted negative case. In other words, true negative occurs when the actual class of an image is negative, and the model correctly predicts it as negative.

Our model has a 95.6% negative predictive value; and although it is considered as an impressive performance, it would help more to understand the reason behind the classification.

Both GradCam and LIME show a more spread-out activity (Fig. 4), which is quite different from positive cases. However, occlusion sensitivity for TN does not look too different from the one of TP, so it cannot be used for comparison.

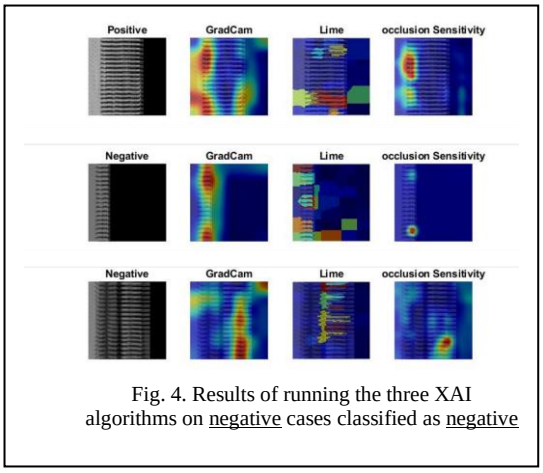

Fig. 4. Results of running the three XAI algorithms on negative cases classified as negative

C. False Negative

A false negative (FN) is an error in binary classification where a negative outcome is predicted when the actual outcome is positive. In other words, a false negative occurs when a model fails to recognize a positive image or when it incorrectly classifies a positive example as negative. False negatives are particularly important in medical diagnoses as

they can lead to the dismissal of sick people when their case ca be critical or even fatal.

All three algorithms show an activity that is less focused than true positives (Fig. 5), which explains the error in classification. It still is different from true negatives.
The images have small areas of high activity (red regions), but also significantly large areas with lesser activity (orange and yellow regions).

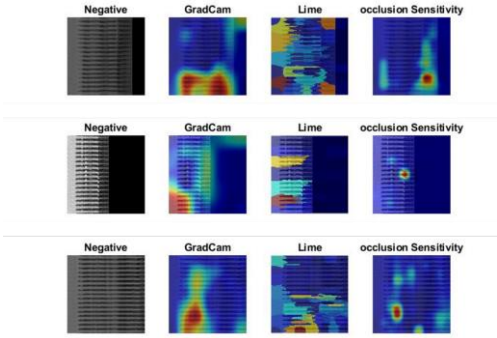

Fig. 5. Results of running the three XAI algorithms on positive cases classified as negative

D. Comparaison between FN and TN

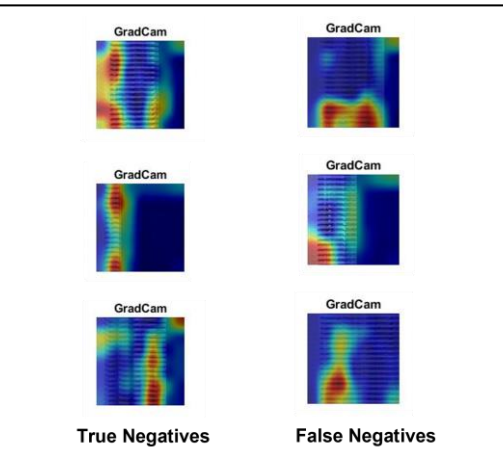

Fig. 6. A comparison between generated heatmaps of negative case classified as negative (left), and positive case classified as negative (right)

As mentioned earlier, false negatives are of critical importance in medical applications, because they refer to patients who are positive but were classified by the model as negative. Their dismissal can lead to fatal complications.

In this experiment, we compared the GradCam analysis of true negative cases and false negative cases. The choice of GradCam was due to it being the only algorithm that showed a significant variation in behavior .

Fig. 6 shows that true negative cases generate heatmaps with a larger surface for high activity (red area) that is well spread vertically. This suggests that a larger region of the image was considered for the final decision of the model. Unlike true positive cases, where only a small region was considered, probably the location of plaque.

As for false negatives, the heatmaps have a smaller localized surface for high activity (red area), with other region with lesser activity (orange and yellow areas). Having the lesser activity can explain WHY the images were falsely classified as negative, a larger area was considered for the final decision.

In conclusion, GradCam shows a tangible difference between false negatives and true negatives, showing significant activity in small areas for false negatives, indicating not only that they were wrongfully classified , but the region for the physician to focus on before making their final decision. Eventually improving the performance of the model.

IV. CONCLUSION

Artificial intelligence has significantly transformed healthcare applications, demonstrating human-level performance on classification tasks in a fraction of the time. However, the opacity in the decision-making process has hindered the trust of many healthcare professionals.

Explainability algorithms serve as a vital bridge to unlock the full potential of AI, augmenting trust alongside superior performance. In this study, we revisited a deep learning model designed for atherosclerosis screening via Coronary CT Angiography. To gain insights into the model's reasoning, we applied three prominent explainable AI algorithms: Gradient-weighted Class Activation Mapping (GradCam), Local Interpretable Model-agnostic Explanations (LIME), and Occlusion sensitivity. The results not only provided additional information regarding classifications but also exhibited a degree of performance enhancement.

Positive cases correctly classified as positive (True Positives, TP) exhibited high focused activity in a specific area, suggesting the presence of plaque. Conversely, positive images erroneously classified as negative (False Negatives, FN) displayed similar concentrated activity, prompting healthcare professionals to reevaluate these regions for a final determination regarding the presence of plaque and atherosclerosis.

Negative cases correctly classified as negative (True Negatives, TN) showcased a more diffuse heatmap, indicating that the model considered a broader image area for its decision-making. GradCam proved to be the most effective algorithm in distinguishing between FN and TN cases.

The primary contribution of this study lies in revealing that the deep learning model primarily relies on the detection of localized high activity, likely representing the position of plaque, thereby rendering the model more trustworthy. Furthermore, the study provides valuable insights by identifying the probable plaque locations in positive cases and, most importantly, offers a means to differentiate true negative cases from false negatives. This augments the

overall performance of the model and reduces the risk of prematurely dismissing patients with underlying conditions.

While the study provides valuable insights, it is limited by the number of samples and should be further generalized in future research. Additionally, it is advisable to test other models with varying performance for comparative analysis.

REFERENCES

[1] H. Lu, Y. Yao, L. Wang, J. Yan, S. Tu, Y. Xie, W. He and others, "Research progress of machine learning and deep learning in intelligent diagnosis of the coronary atherosclerotic heart disease," *Computational and Mathematical Methods in Medicine*, vol. 2022.

[2] C. Molnar, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, Lulu. com, 2020.

[3] S. G. Kim, S. Ryu, H. Kim, K. Jin and J. Cho, "Enhancing the Explainability of AI Models in Nuclear Power Plants with Layer-wise Relevance Propagation," in *Proceedings of the Transactions of the Korean Nuclear Society Virtual Autumn Meeting,*, Jeju, Korea, 2021.

[4] A. I. Suryani, C.-W. Chang, Y.-F. Feng, T.-K. Lin, C.-W. Lin, J.-C. Cheng and C.-Y. Chang, "Lung Tumor Localization and Visualization in Chest X-Ray Images Using Deep Fusion Network and Class Activation Mapping," *IEEE Access*, vol. 10, pp. 124448--124463, 2022.

[5] A. Laidi, M. Ammar, M. E. H. Daho and S. Mahmoudi, "Deep Learning Models for Coronary Atherosclerosis Detection in Coronary CT Angiography," *Current Medical Imaging*, vol. 19, 2023.

[6] A. Rajkomar, J. Dean and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347--1358, 2019.

[7] M. Portegies, P. Koudstaal and M. Ikram, "Cerebrovascular disease," in *Handbook of Clinical Neurology*, F. B. D. F. S. Michael J. Aminoff, Ed., Elsevier, 2016, pp. 239-261.

[8] V. Gupta, M. Bigelow, B. Erdal, L. Prevedello and R. White, "Image dataset for a CNN algorithm development to detect coronary atherosclerosis in coronary CT angiography," *Mendeley Data*, vol. 1, 2019.

[9] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition* , 2016.

[11] M. T. Ribeiro, S. Singh and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization.," *arXiv preprint arXiv:1610.02391*.

[13] "visualizing and understanding convolutional networks.," in *European conference on computer vision*, 2014.