Visual Explainability through Layer Conductance Using the Gradient Ascent Method

Amine Baazzouz*, Jaouad Dabounou

Department of Mathematics Informatics and Engineering Science, Hassan First University of Settat, Settat, Morocco

Research Article

Received: 06-Nov-2024, Manuscript No. GRCS-24-151940; **Editor assigned:** 08-Nov-2024, Pre QC No. GRCS-24-151940 (PQ); **Reviewed:** 22-Nov-2024, QC No. GRCS-24-151940; **Revised:** 08-Dec-2025, Manuscript No. GRCS-24-151940 (R); **Published:** 15-Dec-2025, DOI: 10.4172/2229-371X.16.4.002

***For Correspondence:** Amine B, et al. Department of Mathematics Informatics and Engineering Science, Faculte des Sciences et Techniques, Hassan First University of Settat, Settat, Morocco;
E-mail: a.baazzouz@uhp.ac.ma

Citation: Amine B, et al. Visual Explainability through Layer Conductance Using the Gradient Ascent Method. J Glob Res Comput Sci. 2025;16:002.

ABSTRACT

The significant strides of deep neural network architectures have driven impressive performance gains, but has also introduced greater complexity and opacity within hidden layers, presenting significant challenges for interpreting deep learning models. Various visual explainability techniques have emerged, such as Grad-CAM, Layer-wise Relevance Propagation (LRP), Saliency Maps, and DeepLIFT have been widely used to highlight important regions in an image, attributing these regions to the model's prediction. In this article, we propose a novel approach based on Layer conductance and gradient ascent, typically used to quantify neuron contributions in hidden layers, can also serve as a pathway to visual interpretability. Layer conductance is utilized here in a novel context: Rather than focusing on analyzing hidden units, we extend this approach to examine the impact at the input pixel level, as another gradient-based approach.

Keywords: Neural network; Layer conductance; Visual feature; Explainability; Captum

INTRODUCTION

The evolution of neural networks architectures has led to substantial performance improvements, though it has also amplified the complexity and opacity of their hidden layers, making interpretation a challenging task. To address these challenges, a range of visual explain ability techniques have been developed and refined in the literature, each with unique strengths and applications.

One of these methods is Grad-CAM (Gradient-weighted Class Activation Mapping), has been widely recognized for its ability to produce coarse heat maps that visually highlight important image regions by computing gradients concerning specific classes. Selvaraju et al., [1] demonstrated that Grad-CAM effectively localizes features by leveraging gradients of target classes flowing into the final convolutional layers, making it a preferred method for visualizing decision-relevant regions in classification tasks [2]. Also, DeepLIFT (Deep Learning Important FeaTures), introduced by Shrikumar et al., [3], computes the contribution of each neuron to the model output by comparing activations to a baseline. This approach has shown to be more consistent than simple gradient methods, as it considers both positive and negative contributions, making it well-suited for detailed analysis of the network's behavior. Layer-wise Relevance Propagation (LRP), as described by Bach et al., [4], assigns relevance scores to each pixel by propagating the prediction back through the network. LRP has been especially impactful in its ability to provide pixel-level explanations that make it easier to see which inputs most influence the model's output, supporting transparency in high-stakes applications like medical imaging. Integrated Gradient, first presented by Simonyan et al., [5] compute gradients of the class score concerning input pixels, generating a map that identifies the pixels most affecting the model's decision. Though straightforward, Integrated Gradient provide a valuable initial insight into feature importance, though they may sometimes suffer from noise and lack of interpretive stability in deep models.

The concept of using layer conductance for interpretability through neuron importance in deep learning aligns closely with ongoing research on attribution methods [6]. Layer conductance is traditionally used to identify the importance of specific neurons within hidden layers by integrating activation gradients with neuron contributions, similar to Integrated Gradients (IG) but applied at the neuron level within layers. This approach is especially useful for capturing neuron importance along activation pathways, which can reveal feature interactions deep within the network (as implemented in libraries like Captum by PyTorch).

To address opacity at the pixel level, gradient-based approaches typically backpropagate gradients to highlight regions of interest within the input, allowing us to visualize which parts of an image most impact model decisions. By extending layer conductance to input pixels, our approach adds a novel perspective to this technique, potentially providing deeper insights into how individual pixels contribute across multiple network layers-an aspect that might bridge neuron-level and pixel-level attributions and provide a refined gradient-based explanation at the input level.

Aim and justification

In this article, we introduce a novel application of layer conductance as a pathway to visual interpretability. Traditionally used to quantify contributions of neurons within hidden layers, Layer Conductance is applied in the context of this work to analyze input pixel importance, through the gradient ascent method, allowing to extend its gradient-based analysis from hidden units to input-level interpretation. This approach builds on gradient methods by adding granularity to neuron-based conductance, thus offering an alternative visual explanation for key input regions contributing to the model's prediction.

Layer conductance, when used as a visual explain ability method through gradient ascent, offers several unique advantages for understanding the importance of specific input regions in CNN models. Unlike Grad-CAM or Integrated Gradient, which primarily highlight class-related regions, layer conductance measures the contribution of individual neurons or groups of neurons to the model output. This provides a more nuanced understanding of how each neuron (interpreted as a filter or feature detector) contributes to an image's prediction, which can be particularly useful for detailed visual interpretability. It does not only locate the class in the image, but also clarifies which specific filters, across multiple layers, play a dominant role in the outcome.

Moreover, by tracing gradient ascent back through layers, layer conductance can help identify clusters of neurons that consistently activate for particular input features, even if they are hidden in lower layers. This granularity enables a layered view of feature importance that goes beyond the final convolutional layers, which Grad-CAM focuses on. This approach, therefore, can reveal more about intermediate processing and feature hierarchies within the network.

Traditional gradient-based approaches, like Integrated Gradient, can sometimes produce noisy outputs that are sensitive to minor input changes. Layer conductance, by averaging or aggregating contributions across multiple neurons, can offer a more stable and coherent attribution map, allowing more consistent identification of the most relevant input regions.

Also, Layer conductance is particularly well-suited for models where understanding hidden layers is critical. Grad-CAM and Integrated Gradient are useful for high-level visualization but often overlook details of how intermediate layers interact. By extending layer conductance to analyze pixel-level effects through gradient ascent, we can visualize which input regions strongly activate specific neuron groups, providing a clearer picture of learned high-level and mid-level representations within the network.

Furthermore, in cases where we need to analyze complex or non-discriminative tasks, like concept discovery in multi-object scenes, layer conductance's ability to analyze individual neuron importance makes it valuable. Instead of merely indicating areas associated with a single class, it can show contributions that are consistent across multiple object categories or feature types.

Thus, layer conductance allows for a deep, structured analysis of neuron importance that is less susceptible to noise and interpretable in a way that captures both individual and aggregate neuron behavior, making it a versatile choice for complex interpretability tasks.

Theoretical foundation

Layer conductance: Layer conductance presents a robust alternative to traditional neuron activations for detecting high-level concepts within neural networks [7]. Unlike activations, which measure only the output intensity of neurons, layer conductance quantifies the specific impact each neuron has on the network's overall prediction [8]. This additional layer of analysis offers a precise view of how individual neurons contribute to decision-making within the model, providing valuable insights into the network's inner workings and revealing the underlying concepts that guide its responses [9]. By assessing the influence of neurons across various layers, layer conductance allows for a more comprehensive examination of the internal representations and contributes to enhancing the interpretability and transparency of complex AI models [10].

More specifically, layer conductance estimates feature relevance within each layer, capturing the significance of each hidden neuron in the final prediction. This is achieved by examining connections across neurons, as defined by weights that convey each neuron's influence on subsequent activations.

Let:

- X be the input image
- C^l be the conductance value in layer l
- f be the function representing the neural network up to layer l
- ∇f_X be the gradient of the output with respect to the input.

For a neuron j in layer l, conductance C_j^l is expressed as:

$$C_j^l = A_j^l . \nabla A_j^l$$

Where:

- A_j^l is the activation of neuron j in layer l.
- ∇A_j^l is the gradient of the loss with respect to the activation A_j^l.

A key property of layer conductance is its "Layerwise Conservation Principle", which fully redistributes the network output across all hidden states, capturing both positive and negative values. Another valuable property is "insensitivity," indicating that if altering a hidden unit does not influence the final prediction, the conductance value for that unit will be zero-this reflects the gradient ∇A_j^l within the conductance calculation. These properties make layer conductance a powerful and sensitive method for understanding neural network behavior at a granular level.

Let F_j denote the filter corresponding to image I_i in layer l of the neural network, where j∈{1,2,...,m} and m the number of filters in layer l. The aggregated conductance $C_{i,j}^{agg}$ is defined as the maximum value of $C_{i,j}^l$ across the spatial dimension (height H and width W), and can be expressed as:

$$C_{i,j}^{agg} = max_{x,y}(C_{i,j}^l(x,y))$$

This aggregation of conductance reflects the interpretability of the layer conductance, where the high conductance value indicates the corresponding neuron plays a significant role in contributing to the output.

Gradient ascent for layer conductance: Gradient ascent is a technique where we iteratively adjust input values to maximize (or minimize) a specific target function-in this case, the conductance rather than a simple activation. Conductance offers a more precise measure of each neuron's influence on the model output, so gradient ascent in this context highlights input

regions that contribute strongly to the target neuron's effect on the output. This method can visualize regions in an image that maximally activate a neuron or group of neurons in the network.

Gradient ascent is applied to the input image X to maximize the conductance of a chosen neuron j, updating the image with each iteration as follows:

$$\hat{X} = X + \eta . \nabla(-\max(C_j^l))$$

Where:

- $\hat{X}$ is the updated input image after each gradient ascent step.
- η is the learning rate.
- ∇ represents the gradient with respect to the input image.
- $\max(C_j^l)$ is the maximum conductance value of the selected neuron j.

MATERIALS AND METHODS

The aim of this experiment is to assess and compare visual explain ability techniques to understand how different methods highlight regions of an image that contribute to a model's classification decision. In addition to Grad-CAM, LRP, Integrated Gradient, and DeepLIFT, the Layer Conductance method will be tested using gradient ascent. By analyzing the outputs for four distinct ImageNet classes, this study aims to evaluate the effectiveness of each technique for visual interpretability.

Data and materials: The images are selected from the ImageNet dataset, each from a unique class to ensure that the model's predictions cover a broad range of semantic content. The code is available at https://github.com/AmineBAA/Visual_XAI_Conductance-with-Gradient-Ascent.

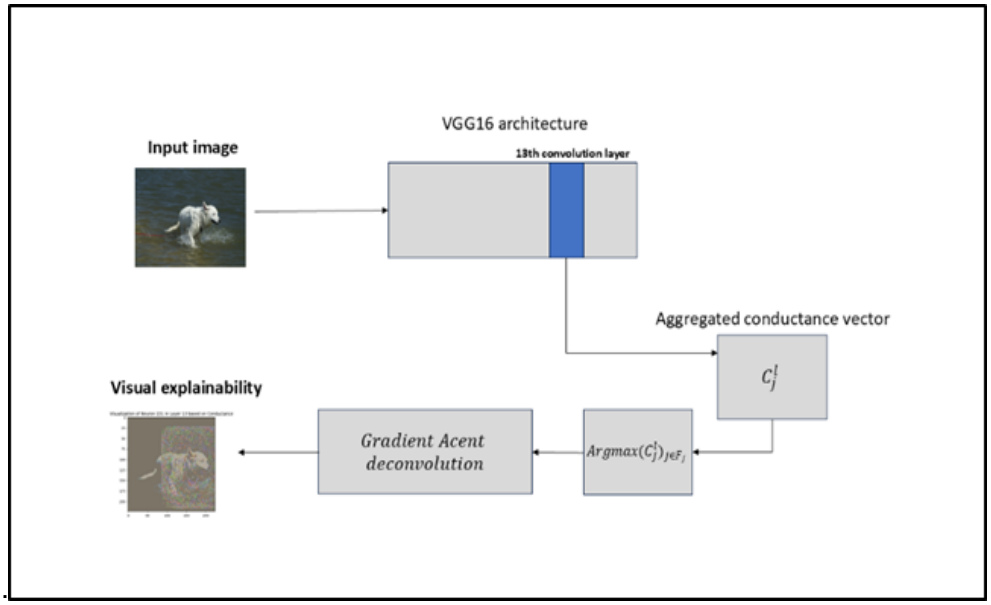
Model and libraries: The neural network used is the pre-trained VGG16 model available in popular deep learning frameworks such as PyTorch or TensorFlow. The library used for explainability methods and for Layer Conductance is Captum.

In VGG16, the layer typically used for Grad-CAM and similar visual explainability methods is the last convolutional layer. This is because Grad-CAM relies on spatial information that is retained in convolutional layers, allowing for a meaningful localization of the important regions of the image. In VGG16, this layer corresponds to layer index 29, which is the final convolutional layer (features.29 in PyTorch) before the fully connected layers.

Using this last convolutional layer is common for methods like Grad-CAM, Layer-wise Relevance Propagation (LRP), and other activation-based attributions, as it has a spatial resolution that can be mapped back onto the input image with sufficient detail, capturing important features without excessive down sampling.

Gradient ascent for conductance: This study explores the detection of high-level concepts within neural networks by employing gradient ascent to visualize learned features, following approaches outlined by Simonyan et al., and Yosinski et al. The initial procedure involves computing the aggregated conductance for images from the ImageNet dataset at last convolutional layer of the VGG16 model, with aggregation performed by extracting the maximum conductance value within each of the 512 filters at this layer. Neurons with high conductance values are selected, as these indicate filters with a strong influence on the model's output. This selection allows for further examination of the specific image regions represented by these filters, giving insight into the features the network associates with high-level concepts. The following Figure 1 illustrates the methodology of this study with steps of proposed approach to visualize important regions in the original image:

Figure 1. Methodology of visual explain ability with layer conductance using gradient ascent.

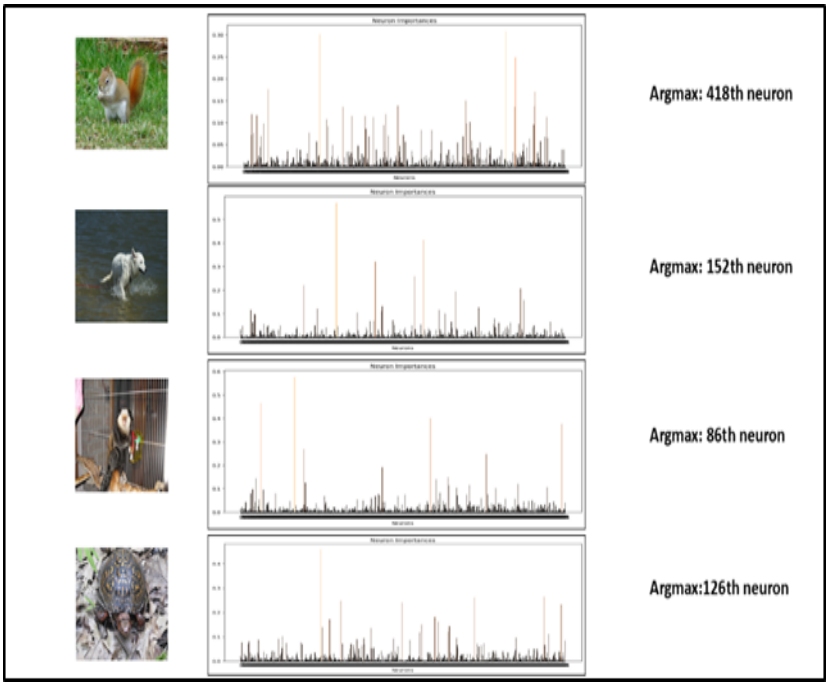


RESULT AND DISCUSSION

Visual explain ability using layer conductance

The experiment suggests that the filter selected from the 13th layer corresponds to the maximum conductance value. The following Figure 1 shows the distribution of aggregated conductance vectors for each image and the rank of neuron with the maximum value (Figure 2).

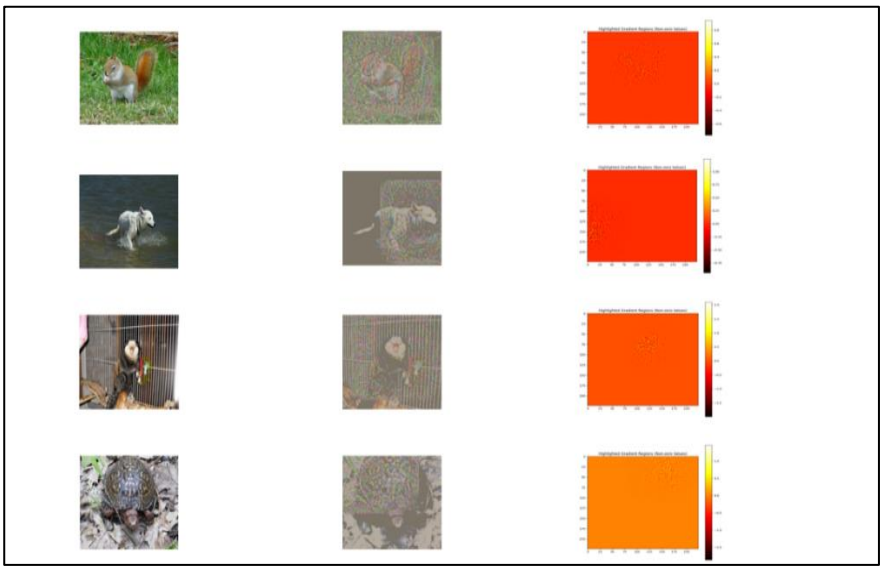
Figure 2. Distribution of aggregated layer conductance for each image at the 13th layer of VGG16.



The next step consists to the deconvolution focusing on the selected neurons using Gradient ascent method to visualize

corresponding regions on the input image. The following Figure 2. Illustrates the impacted regions due to the gradient ascent operation. In the left the original image, in the middle the result of deconvolution corresponding to the selected neurons, and in the right the visualization of impacted regions using only the gradient ∇ with the respect to the input image (Figure 3).

Figure 3. Visualization of selected neuron using gradient ascent method on the original image.



Gradient effect

The Figure 2 illustrates the localization of gradient regions with non-zero values.

Conductance precision stems from the gradient component in its formulation, given by:

$$C_j^l = A_j^l . \nabla A_j^l$$

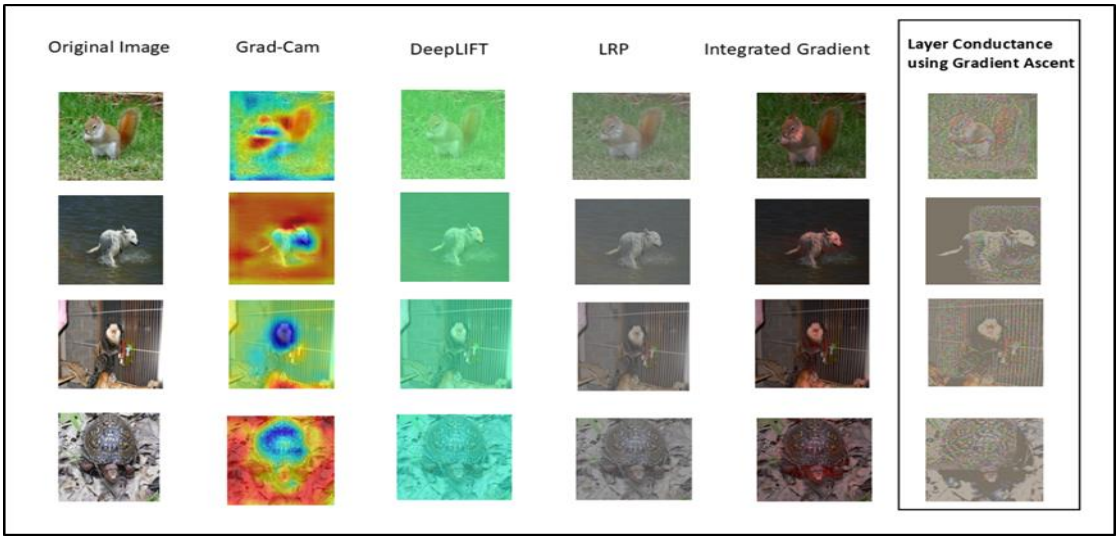
where j denotes the filter index in layer l.

This product of activation and gradient effectively zeroes out certain areas in the input image by nullifying regions in the gradient vector. The figure below highlights the regions with non-zero gradients for the image shown in Figure 1, demonstrating how the gradient focuses on specific parts of the image when deconvolution is performed using conductance.

Comparison between gradient-based methods

The visualization methods are applied to four images to compare between Gradient methods and our proposed approach based on Gradient ascent and layer conductance. The Figure 4. Shows the visualization of impacted regions using different methods:

Figure 4. Comparison of visual explain ability using gradient-based methods from captum library and layer conductance using gradient ascent.



The proposed method of visual explainability using layer conductance and gradient ascent is compared to other interpretability techniques (Grad-CAM, DeepLIFT, LRP, and integrated gradient). Thus, the layer conductance using gradient ascent appears unique among these methods in how it visualizes model attention within specific regions of the input images. It produces subtle, fine-grained patterns with a "noisier" appearance, which suggests it captures intricate details that are otherwise ignored or averaged out in other methods. The conductance-based approach may allow the model to focus on particular neurons or filters that correspond to detailed spatial patterns in the image, leading to an emphasis on textures and smaller-scale features. This could be useful for understanding fine-grained details in cases where the model's subtle activations contribute significantly to the prediction.

Gradient-based techniques:

- **Grad-CAM:** It provides a broad focus on the image areas relevant to the model's output. However, it tends to smooth over finer details, highlighting only the most influential regions. The Grad-CAM visualizations in the figure show bright, focused areas corresponding to general regions of interest.
- **DeepLIFT:** Produces a more muted, slightly transparent overlay of the input image, reflecting its focus on the attribution of model output to individual pixels. It provides a less focused view than Grad-CAM but still highlights areas contributing to the model's decision.
- **LRP (Layer-wise Relevance Propagation):** Similar to DeepLIFT in producing an even, spread-out activation pattern. It captures relevant areas with a moderate level of detail, slightly more focused than DeepLIFT but still less granular than the proposed conductance method.
- **Integrated gradient:** Known for distributing importance across pixels, its output is darker, and it highlights relevant regions diffusely. It captures more generalized regions without specific focus on high-frequency features. The main difference to the proposed approach is that integrated gradient focus on the role of the last convolutional layer in the input image, while layer conductance using gradient ascent focus only on the most important neuron in this layer and visualize its impact on the original image.

The key advantage of layer conductance using gradient ascent is its capacity to highlight detailed regions across the image, capturing a broader range of pixel-level contributions that other methods may overlook. This granularity offers potential insights into the model's behavior on a more microscopic level, which can be critical for certain applications where small features in the input are relevant for classification.

Overall, our proposed method provides a unique advantage in interpretability by capturing finer details and patterns within the input images that the other methods may smooth out. This could be particularly beneficial for tasks that rely on subtle features, making it a valuable addition to the suite of explainability tools for neural networks.

CONCLUSION

Visual explainability offers valuable insights into the concepts that contribute to a model's class predictions, typically through gradient-based methods. In this work, we propose an extended approach, building on layer conductance with gradient ascent to enhance visual interpretability. Our method provides a more detailed visualization of the regions within input images that the model considers important, adding depth to the analysis of feature attention.

In future work, this approach could be expanded by applying it to additional layers within the neural network. This extension would allow us to observe how feature attention evolves across layers, uncovering the roles of individual layers in the model's decision-making process. Such insights would not only enhance our understanding of the neural network's inner workings but also contribute to improved model transparency. By tracking the progression of important regions across layers, we can develop a clearer view of how abstract representations build up in deep networks, ultimately offering a richer, layer-wise perspective on model interpretability.

REFERENCES

1. Selvaraju RR, et al. Grad-cam: Visual explanations from deep networks *via* gradient-based localization. Proc IEEE Int Conf Comput Vis. 2017:618-626.
2. Chattopadhyay P, et al. Grad-CAM++: Improved visual explanations for deep convolutional networks. Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2018:839-847.
3. Shrikumar A, et al. Learning important features through propagating activation differences. Proc Int Conf Mach Learn. 2017:3145-3153.
4. Bach S, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one. 2015;10:e0130140.
5. Simonyan K, et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv. 2013.
6. Sundararajan M, et al. Axiomatic attribution for deep networks. Int Conf Mach Learn. 2017;70:3319-3328.
7. Smilkov D, et al. Smooth Grad: Removing noise by adding noise. Proc 35th Int Conf Mach Learn. 2017;70:4222-4232.
8. Yosinski J, et al. Understanding neural networks through deep visualization. arXiv. 2015:1506. [Google Scholar]
9. Ribeiro MT, et al. Why Should I Trust You? Explaining the Predictions of Any Classifier. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;1135-1144.
10. Kelley A, et al. Interpretation of neural network classifiers through layer conductance. NeurIPS Workshop Interpretable Mach Learn Health Care. 2019:1-8.