1 Introduction

ML has been creating a paradigm shift in medicine in recent years, from basic research (drug discovery and development) to clinical applications(diagnosis and prognosis of diseases such as diabetic retinopathy)[?]. However, as revealed in [?], [?], the ML models are vulnerable to adversarial attacks, where a human-imperceptible manipulation in the input can result in a different predictions or in medical terms a misdiagnosis, which could be utilized for fraudulent interests. This was recently demonstrated by a group of researchers who showed that a carefully calculated perturbation on an image of a benign skin mole that is imperceptible to the human eye can be misclassified as a malignant mole, with 100% confidence [?]. Additionally, the lack of interpretability is another obstacle that prevent ML from being fully trusted. Due to the black-box nature of decision-making process of the deep neural networks, the diagnosis given by the models are not always based on features used by medical professionals. (pathological tumors, inflamed tissues, etc.), making the decision process incomprehensible to humans.

Fortunately, literature on adversarial attacks [?], [?], [?], [?] shows that (restricted to their choice of models in experiments) some widely used deep neural networks can be rendered resistant to the attacks if they are trained on corresponding adversarial examples. Particularly, in [?], Madry et al. claim the projected gradient (PGD) method with random starting point to be the strongest attack utilizing the local first order information (gradients) about the network. In [?], Brendel et al. propose a decision boundary attack (DBA) method that can generate effective adversarial examples solely based on the final decision of the model. They are respectively representatives of white-box methods where we can manipulate the model to calculate gradients and black-box methods where we only have access to the output of the model.

Recently, researchers have demonstrated that the existence of adversarial examples results from the fact that the neural network makes use of some "non robust features" (subtle signals or patterns present in the inputs that are meaningless to humans) for classification [?]. Inspired by this observation, we suppose that through training on adversarial examples, the use of such non robust features will be penalized and the attention of the model should be shifted to the robust features that are equally utilized by humans, which makes the model more interpretable. We wish to verify this hypothesis by training some popular networks for image classification on adversarial examples and employ saliency map methods to visualize the importance the network attaches to each pixel.