# Deep Learning Important Features – Interpretability in Neural Networks

**Project Description:** Neural Networks are becoming increasingly popular. However, adopting them as "black box" machines is facing a challenge in applications where interpretability is essential, such as healthcare.

In this project, we study a newly developed method for decomposing the output prediction of a neural network on a specific input by back-propagating the contributions of all neurons in the network to every feature of the input. In the proposed methodology, the activation of each neuron is compared to its reference activation, and contribution scores are assigned according to the difference. These scores are computed efficiently in a single backward pass. Also, using the difference-from-reference allows the information to propagate even in the case of vanishing gradients, which could prove especially useful in Recurrent Neural Networks. The proposed method avoids placing misleading importance on bias terms, in contrast to gradient-based methods, and reveals dependencies which are otherwise missed. We will also explore how to learn a good reference from the data, and how to best propagate the importance of the scores beyond simply using the gradients.

**Disclaimer and Learning Opportunities.** Students need not have prior knowledge about the concepts mentioned above. They will enhance their programming skills in Python and acquire new knowledge in

- interpretable machine learning;
- gradient-descent methods;
- vanishing gradients;
- Recurrent Neural Networks;
- MNIST dataset;
- Health analytics.

**Duties/Activities:** The intern will run and test different instances of the machine learning code on real data. The code in python will be provided.

**Required Skills:** Python

**Preferred Intern Academic Level:** undergrad / B.Sc.

**Expected Team Size:** 2 students


**Main Mentor:** Dr. Abdelkader Baggag <abaggag@hbku.edu.qa>

**Co-Mentor:** Abdulaziz Al-Homaid <abalhomaid@hbku.edu.qa>