Explainability AI (XAI)
– Understanding the black-box of complex machine learning models

Background:

As we solve more complex problems, the complexity of our models also increases. This means that most machine learning models are doing something under the hood that is so complex that we don't have a clue anymore why they behave the way they do. In other words, we don't know the thought process of our model in predicting something. Understanding the behavior of our machine learning model is becoming very important. Judging the model's performance based on its accuracy or other metrics alone is not sufficient anymore as your model can trick you.

This is where Explainable AI or XAI comes to the rescue. The explanations produced by this concept should help you to understand why the model behaves the way it does. If the model isn't behaving as expected, there's a good chance you did something wrong in the data preparation phase.

Proposal:

We propose to develop several XAI demonstrations across various kinds of machine learning models. Ultimately, we hope to produce a complete no-code product offering that the GWU community working in the field of Image Classification can leverage to evaluate the robustness of built solutions. This tool can be used to evaluate the model beyond the test set.

A comprehensive list of all things we hope to achieve are –

1. Inspection of MLP and LSTM model architectures on demo datasets for image classifications through base PyTorch.
2. Leveraging existing python XAI utilities to provide an in-depth explainer on their internal workings along with a demo for each, to educate users on the right utility for their application. The planned utilities are as follows:
   a. GradCAM
   b. Occlusion Sensitivity
   c. Rise
   d. Deconvnet
   e. Lime
   f. Shap
3. Lastly, develop an entire no-code product offering that will work with TensorFlow and PyTorch frameworks. Users will be able to upload any trained model, Custom or Pre-Trained, along with an observation of their choice to analyze the model's inner working.

Repository & Dataset Source

GitHub Repo: Follow this LINK to view our GitHub repository.