Technical Test: ECG Heartbeat Categorization

# Introduction

The goal of this technical test is for the hiring group to assess your technical skills with a challenge not dissimilar to what you would face on a day-to-day basis. While this exercise requires the completion of several tasks, please feel free to take this as far as you’d like to demonstrate your Data Science skills and set yourself apart from other candidates.

# Dataset Overview

The dataset used in this project is focused on heartbeat classification using deep neural network architectures and explores the capabilities of transfer learning. It consists of electrocardiogram (ECG) signals representing different types of heartbeats. These heartbeats include normal cases as well as those affected by arrhythmias and myocardial infarction. The dataset has been preprocessed and segmented, with each segment corresponding to a single heartbeat.

The Arrhythmia Dataset contains a total of 109,446 samples and is categorized into five classes. The sampling frequency of the ECG signals is 125 Hz, which indicates that the signals were recorded at a rate of 125 samples per second. The dataset is sourced from Physionet's MIT-BIH Arrhythmia Dataset, a well-known resource in the field of biomedical signal processing. The classes in this dataset are represented by the following labels: 'N' (0), 'S' (1), 'V' (2), 'F' (3), and 'Q' (4). These labels correspond to different types of arrhythmias and myocardial infarction cases.

You can download the dataset from [kaggle](https://www.kaggle.com/datasets/shayanfazeli/heartbeat?resource=download&select=mitbih_train.csv).

# Test Objective

Leveraging the dataset mentioned above, the overall goal is to create a classification model and data pipeline that will classify a segment of ECG as any one of the provided heartbeat classification labels. Please structure your test base on the tasks listed below. A complete solution is more important than model performance.

# Structure of Submission

Follow the [cookiecutter data science structure](https://drivendata.github.io/cookiecutter-data-science/) for organizing your project, please include your explanations/documentation as markdown files within the submission

# Use of ChatGPT or Other Aides

We understand that most people are using aides in the development practices these days, and we understand that. We weight the explanations higher that submitted code and will ask questions about your solution during the interview phase.

# Technical details

## Task 1: Data Processing

* Perform an Exploratory Data Analysis (EDA) on the ECG Heartbeat Categorization Dataset. Provide insights into the distribution of classes, statistical properties of the features, and any other relevant observations.
* Apply data augmentation techniques to enhance the diversity of the dataset. Explain the rationale behind the chosen techniques and how they contribute to improving the model's performance.
* (If necessary) Perform feature engineering to extract meaningful information from the raw ECG data. Select and justify the features you believe are most relevant for the heartbeat categorization task.
* Handle missing data by applying appropriate data imputation techniques. Discuss the impact of missing data on the model's performance and the reasons behind your chosen imputation method.

## Task 2: Model Training and Fine-tuning

* Choose a suitable model architecture for the ECG heartbeat categorization task. Explain your choice and discuss the advantages and potential challenges associated with the selected model.
* (Optional) Perform hyperparameter tuning to optimize the model's performance. Clearly state the hyperparameters you have selected for tuning and describe the methodology used. You are free to use any tool(s) you deem necessary (MLFlow, Hydra…)
* Address issues related to overfitting and underfitting during the training process. Explain the techniques you have employed to cope with these challenges.
* Implement early stopping and training callbacks to improve the training efficiency and prevent overfitting. Discuss the rationale behind their usage.
* Evaluate the trained model using appropriate evaluation metrics. Provide a detailed analysis of the model's performance and discuss any limitations or potential areas for improvement. You are free to use any tool(s) you deem necessary (MLFlow, DVC, Hydra…)

## Task 3: Testing Holdout Set

* Create a holdout set from the ECG Heartbeat Categorization Dataset to simulate data shifts and performance degradation. The holdout set is different from the validation and testing set.
* Apply the trained model on the holdout set and evaluate its performance. Analyze any differences in performance compared to the training/validation/test results and discuss potential causes.

## Task 4: Deployment Strategies

* Create a model inference pipeline that will take a segment of data and generate a prediction.
* Implement or propose a deployment strategy for a production-ready solution based on the trained model. Discuss the considerations, challenges, and best practices involved in deploying a machine learning model in a real-world scenario. You are free to use any tool(s) you deem necessary.
* Describe how you would handle model versioning, scalability, and monitoring in the deployed system. Discuss any potential risks and mitigation strategies.

Please document your work in a well-organized manner, including clear explanations, code snippets, visualizations, and any other relevant materials. Feel free to include additional information or experiments to showcase your skills and expertise. You have one week to complete the test.

# Submitting your work

To submit your project, please follow the steps below:

* Push your work into a private GitHub repository: Create a **private** repository on GitHub and push all your project files, code, documentation (markdown files), and any other relevant materials into the repository. Ensure that your repository is well-organized and includes clear instructions on how to run your code and reproduce your results.
* Add the SickKids reviewers as a viewer by GitHub tag: "@I-Akrout" and “@bgreer101”.

Once you have completed these steps, please reply to this email and let us know you’re complete and provide the necessary details for accessing your repository.