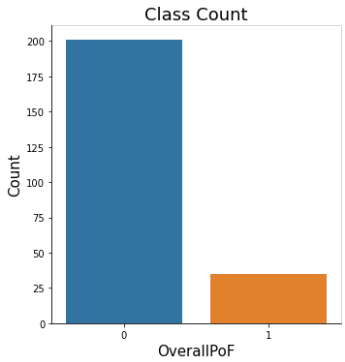
Exploratory Data Analysis

EDA →1.data imbalanced → methods + large dataset → SMOTE + Generative d
→ 2. Multicollinearity (corelated frequencies) → feature selection

First figure 1 makes it abundantly evident that there were much more patients who passed the overall test (Overall=0) than there were patients who failed it (anomaly). Consisted of a small number of abnormal groups and many normal groups, there was a definite 'class imbalance problem' [ref]. As there is a need to generate data to address this issue since real-world medical data frequently contains anomalous data with few abnormalities, our team has also used techniques like VAE to finish the exploration data generation stage.



There is high correlation of absorbance for nearby frequencies with correlation reducing with distance (Figure 1, right). Correlation is a special case of multicollinearity, and high correlation implies multicollinearity. If keep all frequencies would pose Multicollinearity when utilising machine learning models, eventually leading to unreliable results. Therefore, it was necessary to perform a feature selection of the frequencies, selecting the 20 most meaningful frequencies from the 125 frequencies from 226Hz to 8000Hz, which would also reflect the patient's WBA profile. This move also alleviated the pressure of data collection and model training.