1. Convert the PDF pages to images.
2. Separate tables and text and remove noise (header, footer, table descriptions, etc.).
3. Use a pretrained OCR model to read the **tables** and the **Condition Text** from the images (this can be done in parallel).
   1. Fine-tune the OCR model on the dataset (*only during training*) (not implemented in the POC).
      1. Evaluation metrics –
         1. Offline – Accuracy on test data.
         2. Online – Number of words wrongly identified.
   2. For the text, use the model to identify bolded and underlined text and list items (not implemented in the POC).
   3. Use text properties (bolded, underlined, indentation of list item) to identify the hierarchy of the conditions (**Condition Number**) (not implemented in the POC).
4. Train a classification model to predict the Classification (DSC, REQ, RAE) given the text (*only during training*).
   1. Evaluation metrics –
      1. Offline – F1 score, Area under ROC Curve for test data.
      2. Online – Accuracy with respect to manual labelling.
5. Use the classification model identify the **Classification**.
6. **Requirement Description** –
   1. None for DSC and RAE.
   2. For REQ – same as the text if not under an RAE, and otherwise the concatenation of the texts for the parent RAE and the requirement.
7. **Requirement Name** –
   1. None for DSC and RAE.
   2. For REQ – same as the Condition Number.
8. **Referable Conditions** –
   1. None for DSC.
   2. For REQ – All conditions in the hierarchy up to the current element. For e.g., for 1.6.C.i. – 1., 1.6., 1.6.C., 1.6.C.i.
   3. For RAE – All conditions in the hierarchy up to the current element as well as all sub-headings. For e.g., for 1.6.D. – 1., 1.6., 1.6.D., 1.6.D.i., 1.6.D.ii., 1.6.D.iii., 1.6.D.iv., 1.6.D.v., 1.6.D.vi.