Assignment - Machine Learning

Objective:

The objective of this assignment is to build an NLP solution for the provided dataset. The dataset consists of scanned documents from an archive.

Requirements:

1. The dataset can be downloaded from https://www.sec.gov/Archives/edgar/vprr/index.html Choose all files in directories starting with 00 or 01. These contain scanned documents - mostly different kinds of regulatory forms and other documents.
2. Create a model for classifying a document into one of the form types or if the document isn't a form, then a category called "Other"
3. Check the accuracy of your model.
4. Do not use any cloud services for any part of the task required for the classification.
5. Do not copy the solution from any existing repositories. Submit your own code.

Outcomes:

1. A model that can categorize a form with as high an accuracy as possible
2. A documentation of your approach along with reasoning for your model choices
3. Model evaluation techniques used for classification.
4. Balance between accuracy and speed for simple vs complex models
5. Accuracy and Speed metrics.
6. You can commit this code and document to a public or private repo on a service like Github or Gitlab and share with us.
7. Usually this is a task that will take about a weekend for someone familiar with the techniques involved. In case you think it will take you more time, please let us know the time required and we can adjust accordingly.