**Document Classifier using Deep Learning**

# Table of Content

Table of Content 2

Problem Statement 3

Required Input 3

Format of Input 3

Directory Structure 3

Expected output 3

Model Architecture 4

# Problem Statement

In BFSI sector, banks and insurance deal with various documents requested from the customers banking and insurance services to be provided. The documents maybe scanned or handwritten or may vary in the format in which it is provided to the banks and insurers. It is a tedious task for people to manually segregate or classify the documents manually. In this era of digital world, automating the classification of documents would make it easier to reduce manual effort in classifying document. Automation shall be achieved by creating a Deep Learning based AI model that will predict the document class name based on the type.

# Required Input

The required input would be images and pdf files.

# Format of Input

* Images of different extensions like JPG, JPEG, PNG, TIFF etc.
* First page of PDF file would be converted to image and stored in PNG format.

# Directory Structure

Dataset

|\_\_ raw -- (Mix of both PDF and Images)

|\_\_ class1

|\_\_ class2

|\_\_ class3

|\_\_ Images

|\_\_ class1

|\_\_ class2

|\_\_ class3

Pdf\_to\_image.py

Image\_preprocessing.py

Data\_loader.py

Model.py

Accuracy\_loss.py

Prediction.py

Testing.py

# Expected output

Expected output will be class label name that will be predicted by the AI model along with a confidence score that shall range between 0 and 1.

# Model Architecture

* The AI model shall consist of stacked CNN model of two transfer learning based CNN architecture:
  + VGG – 16
  + ResNet – 50.
* The include\_top parameter of model initialisation shall be False to extract the bottleneck features
* The dataset shall consist of three classes.
* Train Test Split will be done with ratio of 70:30 with a random\_state of 42 to ensure same train\_test split is maintained in multiple runs.