Lohit Hostel

Automated Research paper Categorization

Problem Statement

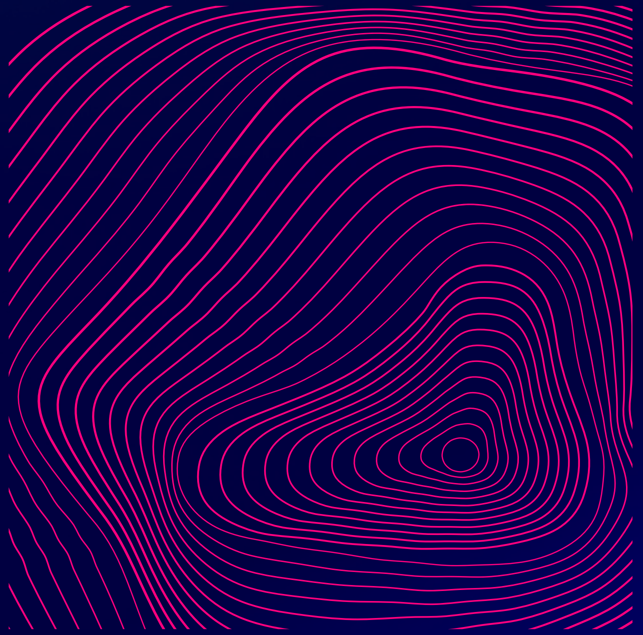
Building an Automated Research paper categorizer using Machine Learning and Deep Learning techniques.

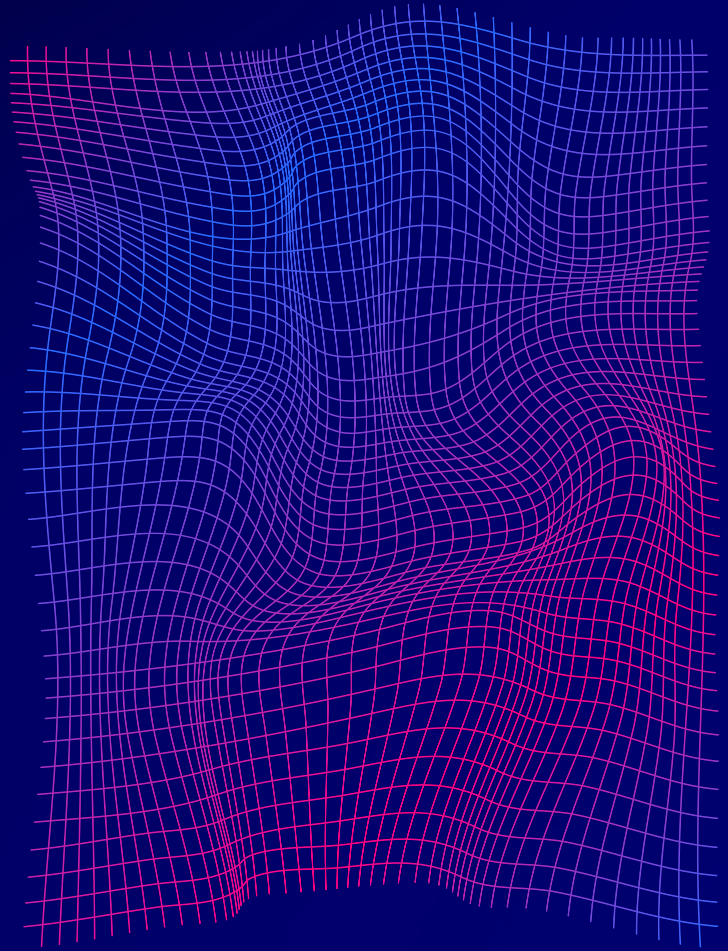
DATASET DESCRIPTION:

- The train and test datasets consist of multiple research papers.
- The features preset in the dataset are:
    - Title of research paper
    - Abstract of research paper
    - Categories to which the paper belongs to
    - Number of rows in train dataset is 51210
    - Number of rows in test dataset is 10974
    - Number of categories is 57

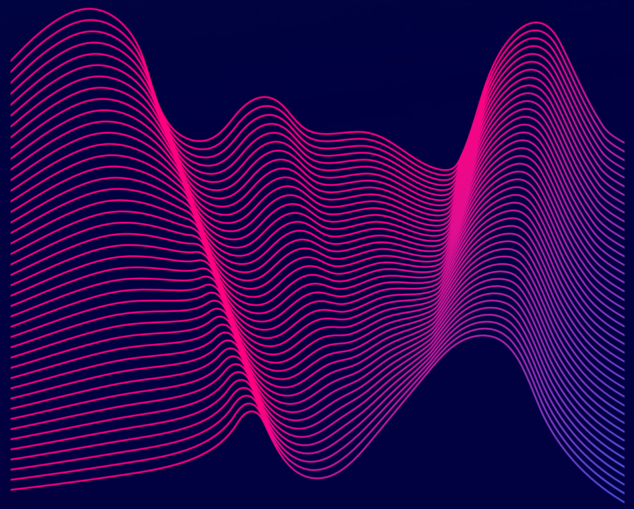
A peak at the dataset

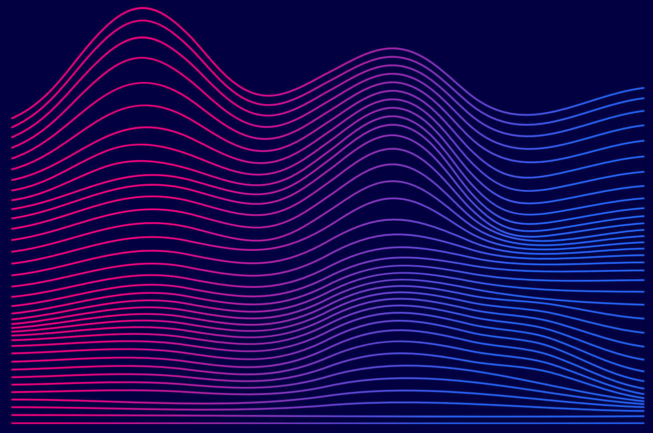
Title	Categories
Large deviations for Wishart processes	['math.PR']
Slicer Networks	['eess.IV', 'cs.AI', 'cs.CV']
New symmetry in nucleotide sequences	['q-bio.GN', 'q-bio.BM']
Modeling Credit Risk with Partial Information	['math.PR', 'q-fin.RM']
A Semantic Grid Oriented to E-Tourism	['cs.DC']

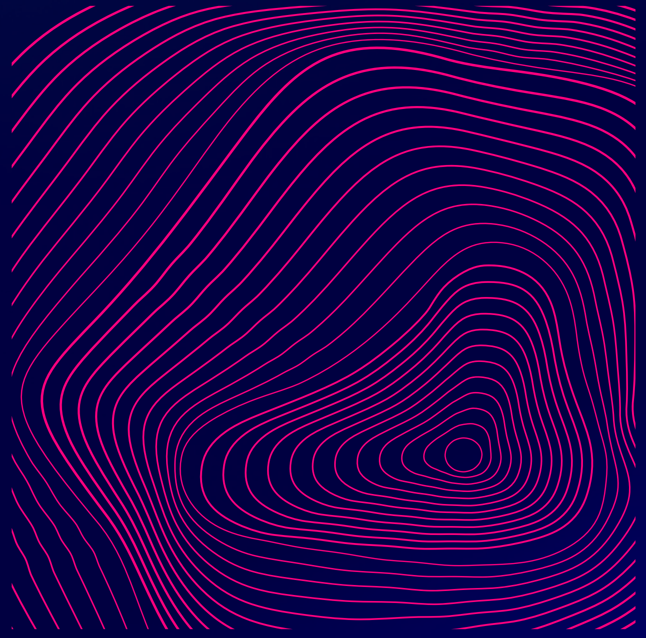


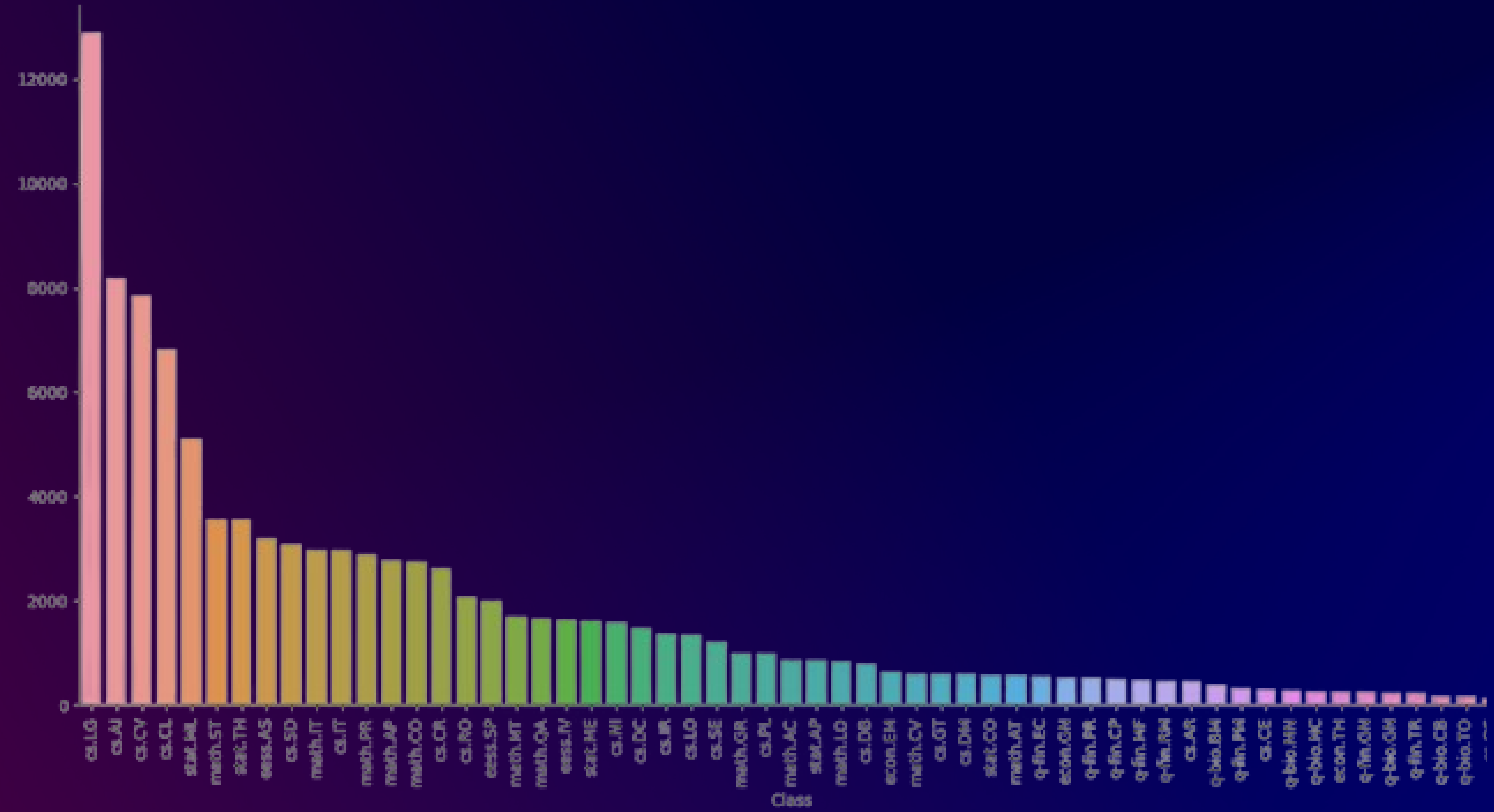


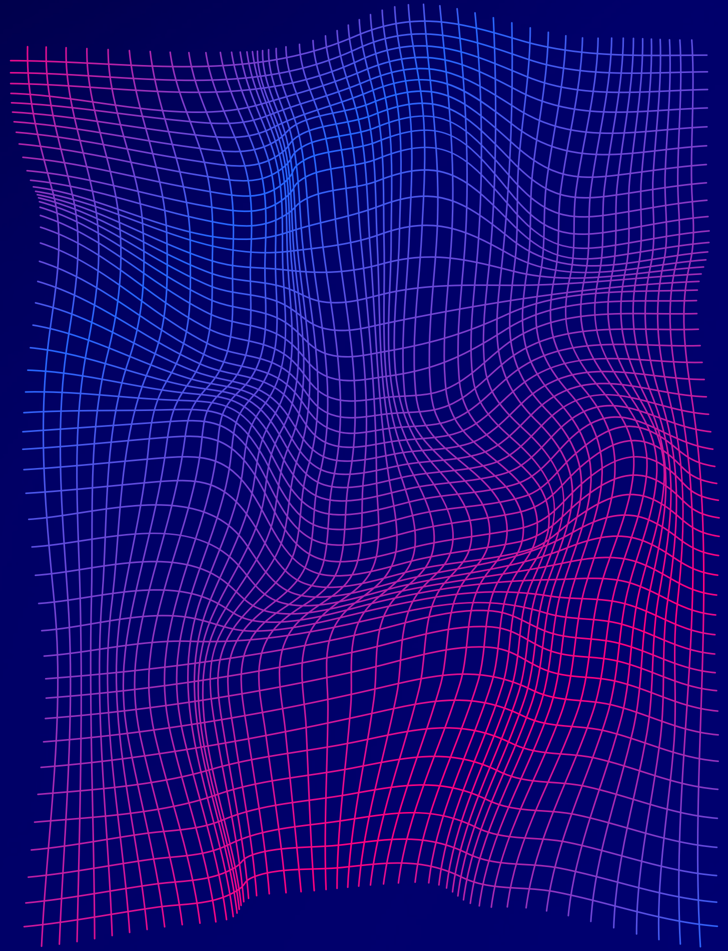
Class Imbalance











Data Preprocessing

- Class Weighting:
  - Assigned higher weights to minority classes during training to penalize misclassifications in these classes more heavily.
- Resampling:
  - Resampled the dataset to balance the class distribution. This involved randomly selecting samples from the majority class or creating synthetic samples for minority classes.

Text Preprocessing Summary

1 Merged the ‘Abstract’ and the ‘Title’ columns into a column called ‘Context’ which is finally used for Prediction.

2 This was followed by decontraction of some word like won’t ,can’t to ‘will not’ and ‘can not respectively’

3 We removed all the punctuation and stop word from the test.

4 Then we proceeded with stemming all the word to their root word . (Like ‘Happier’ to ‘Happy’ , ‘Programming’ to ‘Program’ and so on)

5 Then we formed a vocabulary using all the word present in the text.

6 Finally each row in the dataset was converted into a vector where each word in text was replaced with its position number in the vocab

Approaches

DL based approaches

CNN and Bi-LSTM

Gave a Public F1 score of 0.56

- Used an Embedding Layer
- Fed the embedding outputs to a 1D Conv layer
- Used a bidirectional LSTM layer followed by feeding it into a max pooling layer
- Feeded the outputs to subsequent dense layers
- Used an output layer with a sigmoid activation

Bert transformer model

Gave a Public F1 score of 0.61

- Used a pre-trained Bert model for multi-label classification
- Fine-tuned it on the train dataset

Dense Neural Network

Gave a Public F1 score of 0.65

- Used 3 dense layers with BatchNormalization and Dropout
- Used swish activation for initial layers with leaky relu for the last layer
- Used adam optimizer for model training

ML Approaches

XGBoost Classifier

Gave a Public F1 score of 0.54

- Used XGBoost with calibrated classifier CV
- Created a pipeline for each category using Tfidf Vectorizer and OnevsRestClassifier for multi-label classification

Support Vector machine

Gave a Public F1 score of 0.62

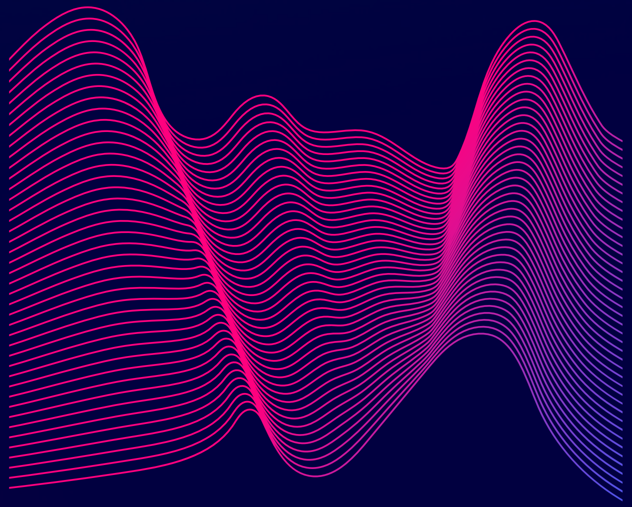
- Used SVM with calibrated classifier CV
- Created a pipeline for each category using Tfidf Vectorizer and OnevsRestClassifier for multi-label classification

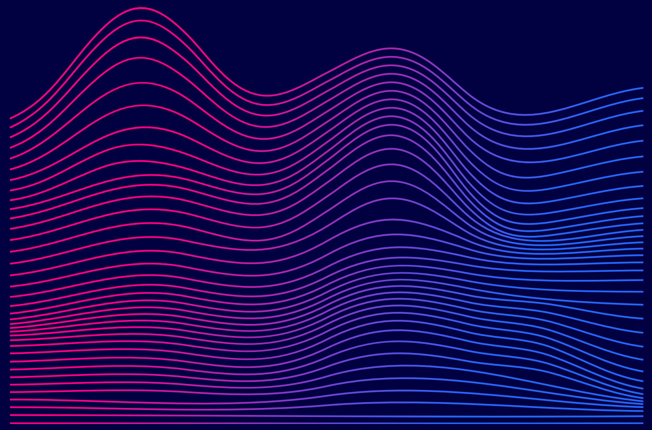
Final approach

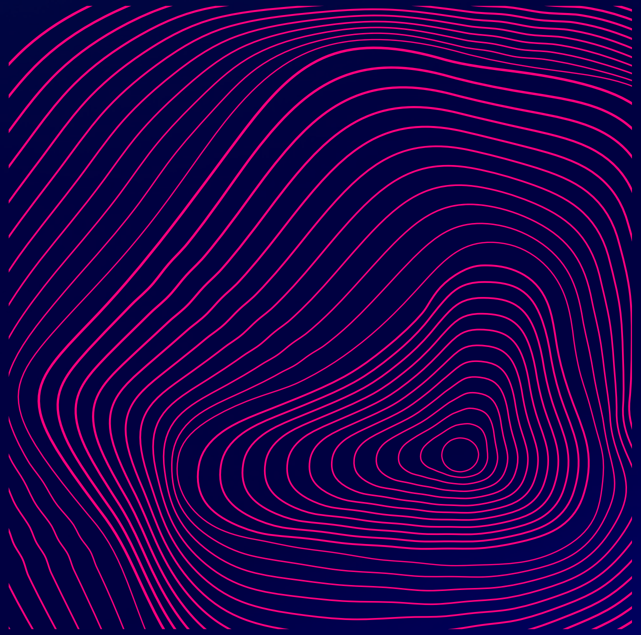
Dense Neural Network

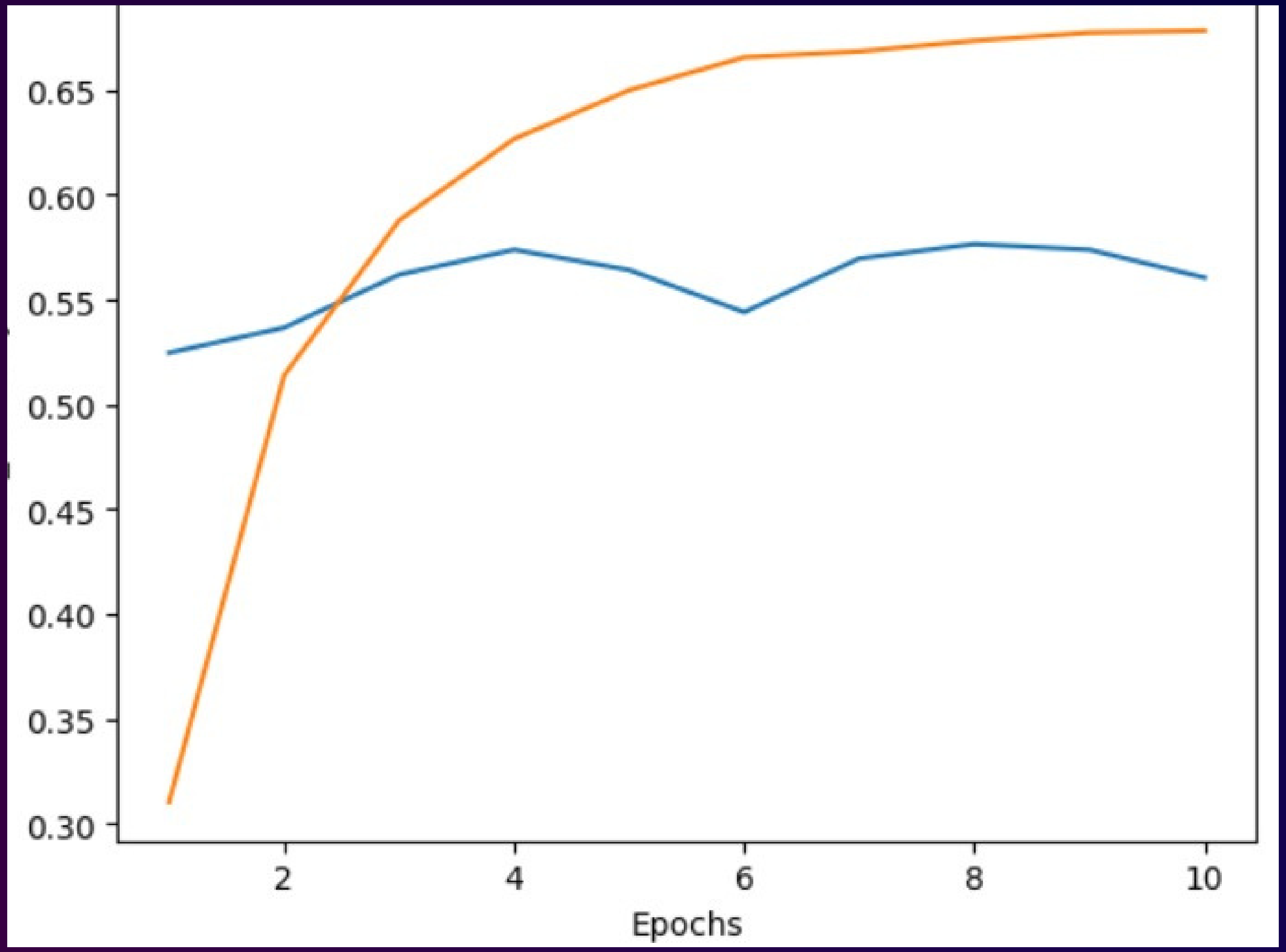
- Used 3 dense layers with BatchNormalization and Dropout
- Used swish activation for initial layers with leaky relu for the last layer
- Used adam optimizer for model training

Training Logs

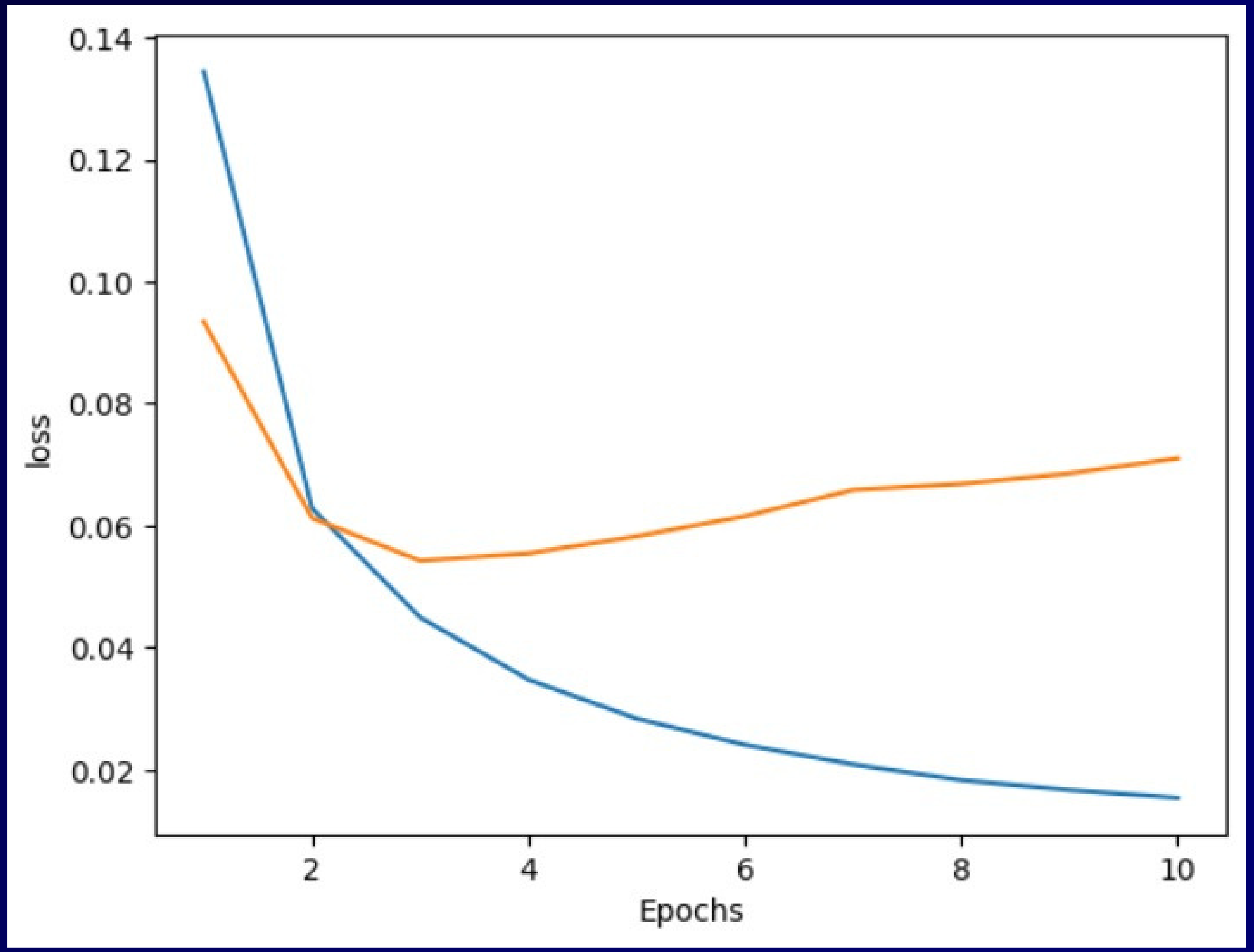








Training accuracy vs Validation Accuracy



Training loss vs Validation loss

Thank You