Winter 2025 CIS*6180/Data*6300 Analysis of Big Data
Assignment #2
Total Marks: 15; Due date: March 28, 2025 (11.59 pm)

What you need to submit:
A single zip file that contains the following:

1. A PDF with analysis, answers, and visualizations.
2. Any CSV files required for results
3. Your code

Programming language: Use as per the course outline.

In this assignment, you will:

- Train a deep learning model to classify high-dimensional data.
- Experiment with different neural network architectures.
- Evaluate and visualize model performance.

DataSet (*same as assignment #1*): You will use the provided dataset ("dp_with_labels.csv"). It contains 1550 data points, each represented as a feature vector of 4096 features. The first column contains datapoint labels (DP0001, DP0002, ..., DP1550). The remaining 4096 columns are numerical features. So, each row contains a data point label (column 1) and its respective feature vector (columns 2 - 4097). Here is the class distribution: *Class#1 (First 425 points), Class#2 (Next 400 points), Class#3 (Next 375 points), Class#4 (Next 350 points)*.

Task 1 (9 marks): Train a deep learning model to classify data points based on their labels.
Steps:

1. (1 mark) Load and split the dataset into training (80%) and testing (20%) sets.
2. (3 marks) Implement a feedforward neural network using the following configuration:
   a. *Input layer:* Accepts 4096 features.
   b. *Hidden layers:* Use ReLU activation, and apply techniques such as dropout or batch normalization.
   c. *Output layer:* Uses softmax activation (multi-class classification).
3. (1 marks) Train the model using an appropriate loss function (e.g., cross-entropy for classification).
4. (1 mark) Evaluate the model on the test set.
5. (3 marks) Include the following in the PDF:
   a. Visualize the loss and accuracy curves during training.
   b. How does your model perform? Briefly discuss accuracy, loss, and any overfitting issues.
   c. What hyperparameters did you choose (batch size, learning rate, optimizer, etc.)?

Task 2 (6 marks): Hyperparameter Tuning
Steps:

1. (2 marks) Experiment with at least two more different architectures by modifying a combination of Number of hidden layers, Number of neurons per layer, Activation functions, and Dropout rates.
2. (2 marks) Compare their performance with the original model using a table showing:
   Training Accuracy | Test Accuracy | Observations
3. (2 marks) Include the following in the PDF:
   a. Which architecture performed best? Why?
   b. How does increasing/decreasing model complexity affect results?