**Deep Learning Assignment 1**

**Soil Type Classification Using PSA Data**

### **Overview**

In this assignment, you will develop and train a deep neural network to classify soil types (Hard, Transitional, or Soft) based on real-world seismic data from stations across Mexico. You will:

1. **Explore** the provided PSA dataset and understand the features.
2. **Preprocess** the data to ensure suitability for a deep learning model.
3. **Build and train** a neural network to classify soil type.
4. **Evaluate** the model’s performance using metrics such as accuracy, precision, recall, and F1-score.
5. **Discuss** potential improvements and real-world implications of your model.

### **Task A: Data Exploration and Preprocessing**

1. **Load the Dataset**
   * Use pandas (or a similar library) to read in the dataset.
   * Inspect the first few rows to understand the data structure.
2. **Exploratory Data Analysis (EDA)**
   * Print summary statistics for numerical columns.
   * Check for missing values. Decide on a strategy to handle them (imputation or removal).
   * Visualize class distribution (Hard, Transitional, Soft).
     + A bar plot or pie chart showing how many samples belong to each soil category.
   * (Optional) Plot histograms or boxplots for key features like magnitude, depth, or max accelerations to see if they differ by soil type.
3. **Feature Engineering**
   * **Option 1 (Reduced Feature Set):** Use summary statistics of the Spectral Ratio (mean, max, min, median, std) + basic earthquake measurements.
   * **Option 2 (Full Spectral Ratio):** Use all frequency components (e.g., 1,000 frequency bins) plus the CEM features.
   * **Option 3 (PCA or Other Dimensionality Reduction):** Apply Principal Component Analysis (PCA) on the high-dimensional SR data to reduce to a smaller number of components while preserving most variance.
4. **Data Splitting**
   * Divide the data into **training**, **validation**, and **test** sets.
   * For instance, 70% training, 15% validation, 15% test.
   * Ensure stratification so that each split has a proportional representation of Hard, Transitional, and Soft.
5. **Data Scaling**
   * If you choose to include high-dimensional numeric features, consider applying standardization (z-score) or min-max normalization, especially beneficial for neural networks.
6. **Label Encoding**
   * Convert the soil categories (Hard, Transitional, Soft) into numeric labels (e.g., 0, 1, 2).

**Task B: Model Architecture & Implementation**

* Implement a deep neural network using either Keras (TensorFlow). Evaluate model on test set, create classification metrics, confusion matrix and plot training curves.

**Task C: Interpretation**

**Summarize how the model performed:**

* Did the model achieve an accuracy close to or exceeding the 89% benchmark from prior machine learning methods (e.g., XGBoost in the paper)?
* How do the results compare using different feature sets (e.g., PCA vs. full SR features)?

**Challenges & Limitations**

* Discuss the **class imbalance** (more samples in Hard vs. fewer in Soft/Transitional) and how it impacts performance.
* Consider data consistency issues across stations or events (e.g., missing frequencies or sensor malfunctions).
* Explain potential improvement.
* Briefly describe how an automated method for soil type classification can benefit seismic hazard assessment and earthquake engineering (e.g., cost savings, rapid site analysis).