Assignment 4

Due Thursday, February 29th before noon (California time)

Dynamic changes in gene expression, orchestrated by complex internal and external signaling networks, are crucial in determining cellular states and fates. Despite the variability observed at the single-cell level, Raj and Van Oudenaarden (Cell, 2008) highlight that predictable patterns of gene expression emerge over time or across cell populations. In this assignment, you will develop a Long Short-Term Memory (LSTM) network to analyze the temporal dynamics of gene expression data, aiming to capture these temporal dependencies. You will then compare its performance with that of fully connected neural networks (FCNN). This task focuses on the challenge of classifying and predicting gene expression dynamics across various groups of genes.

1. **Build a Fully Connected Neural Network (FCNN) for Gene Expression Classification [2pt]:** Begin by loading the dynamic gene expression data from *gene_expression_data.csv*, which comprises 1000 samples, each featuring 10 genes across 20 time steps with normalized expression levels. Your task is to develop an FCNN that can classify gene expression dynamics into two distinct classes as indicated by the *'outcome'* column in the dataset—'1' for *upregulation* and '0' for *downregulation*. After loading the dataset, divide it into training and testing sets, maintaining a test size of 20%. Design your FCNN to include two dense layers and embark on the training process for a duration of 30 epochs.

2. **Model Performance Evaluation and Analysis [2pt]:** Evaluate your FCNN's performance by examining the training and testing *accuracy* and *loss*. Use *history* and *plots* functions to capture and visualize these metrics and provide an analysis of your findings. Discuss any observed issues with the training or testing results and suggest applicable techniques from our coursework to address them. Justify your choice of techniques.

3. **LSTM Model for Temporal Gene Expression Classification [2pt]:** Shift focus to predicting gene expression dynamics by developing an LSTM model that captures temporal dependencies. Load and preprocess the dataset to match the LSTM's expected input shape (num_samples, num_time_steps, num_genes), basing your reshaping on the provided column names, and keep a test size of 20%. Design the LSTM for binary classification, embarking on a training process for a duration of 30 epochs with the goal of surpassing a test accuracy of 0.8. You can start with LSTM's units=50.

4. **LSTM Model Evaluation and FCNN Comparison [2pt]:** Post-training, evaluate your LSTM model as you do for the Q2. Provide an analysis of the LSTM's results, comparing its performance against the FCNN. Conclude by discussing the implications of your findings and draw conclusions on the effectiveness of LSTM models in capturing temporal gene expression dynamics versus FCNN.