**CS 6304 – Deep Learning**

**DATA:** The data given in two CSV files (training data: **dm3.kc167.tads.train.csv** and testing data: **dm3.kc167.tads.test.csv**) represent two classes of DNA sequences (class 0 and class 1). Sequences with class label 1 interact with a particular protein, and sequences with class label 0 do not. One DNA sequence is a long sequence of four letters (A or a, T or t, G or g, and C or C). Some sequences may contain ‘N’ which means is missing letter. All DNA sequences in the dataset have the same length. The sequences and class labels are in the first and second columns, respectively.

**TENSOR PREPARATION:** Convert each letter to a vector of length 4 using “one-hot encoding” as below.

* ‘A’ or ‘a becomes [1, 0, 0, 0]
* ‘T’ or ‘t’ becomes [0, 1, 0, 0]
* ‘G’ or ‘g’ becomes [0, 0, 1, 0]
* ‘C’ or ‘c’ becomes [0, 0, 0, 1]
* ‘N’ or ‘n’ becomes [0,0,0,0]

That means, the DNA sequences data set should be numerically converted to a 3D tensor of shape (N, L, 4), where N, L are the number of sequences and the length of a sequence, respectively.

**TASKS:** Given below are 13 different model architectures that can be trained on the given dataset. Your goal is to train at least 7 models from the following list. You are free to design your architecture. Report the accuracies on the test set for all the models designed. Aim for at least 85% accuracy. More than 90% accuracy is possible.

A table with text and numbers

Description automatically generated

**Table 1: List of model architectures**

Because of time and computational limitations, you are not required to perform hyperparameter optimization. However, you are given some suggestions for values of some important hyperparameters.

* Number of kernels = 16, 32, 64, 96.
* Length of kernel = 9, 15.
* Learning rate of the optimizer: 0.001, 0.005, 0.01.
* You should use the ADAM optimizer.
* Number of LSTM neurons in a block: 10, 20, 30, 40.
* Number of Dense Layers: 1, 2,3, 4.
* All pooling layers are local max pooling layers.
* Include dropout in CNN, LSTM, and Dense layers.

**Grading Rubrics:**

* 10 points: Transferring the CSV files to proper data representation for model input and training.
* Setting up the right architecture of each model (adding the right layers). Follows proper training procedures for each model, such as “early stopping criteria”, “validation split”, and “dropout” to prevent overfitting.
  + 10 points per model (total 70 points).
* 10 points: Depending on the best accuracy you get, you can get up to a maximum of 10 points.