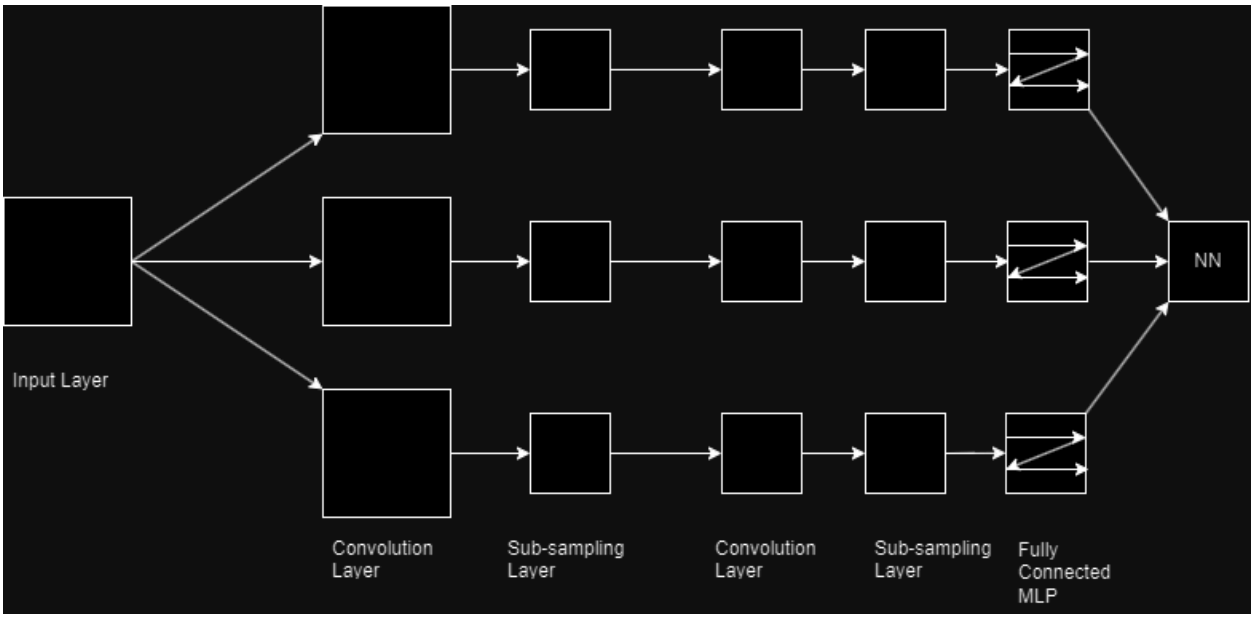
Question 1 (15 points):

1) Please explain the technique of Gaussian Mixture and how it is used for anomaly detection. (5 points)
    a) The Gaussian Mixture Model is a probabilistic model generated using the assumption that instances are generated from a mixture of several Gaussian distributions with unknown parameters. The Expectation-Maximization algorithm can be used to estimate the parameters of the GMM. This involves several steps:
        i) Initialize the hypotheses, assuming only the means of the distributions are known (but not the parameters).
        ii) Calculate the expected value for each latent variable, assuming the current hypotheses hold.
        iii) Calculate a new maximum likelihood hypothesis, assuming each latent variable takes the value calculated in Step 1. Then, repeat this, replacing the current hypothesis with the next hypothesis.
        iv) Finally, we find a model that minimizes a theoretical information criterion, using either the Bayesian Information Criterion, or the Akaike Information Criterion.
    b) The advantage of the Gaussian Mixture model is that it works very well with clusters with different ellipsoidal shapes, sizes, densities, and orientations, while K-Means does not.
    c) There are two ways that the Gaussian Mixture Model can be used for anomaly detection:
        i) The first is distancing. This involves calculating the distance between the new data points and the Gaussian distributions' centroids in the model. Any data point that is far from the centroid can be considered an anomaly.
        ii) The second is thresholding. This involves setting a threshold on each data point's probability, and if that data point's probability is below the threshold, then it is an anomaly.

2) Please draw the diagram of Convolutional Neural Networks (CNN). Then explain the functionality of each layer of CNN. Name several latest algorithms of CNN (e.g., AlexNet, etc.). (5 points)

a)



b) Functionality of each layer:
   i) The convolutional layer changes the image using trainable filters. The filters identify various features of the image, such as corners or edges, which then allow the filters to create a collection of feature maps. These maps then show which features are in the image.
   ii) The sub-sampling layer takes the feature maps from the convolutional layer and down samples them. This increases the Neural Network's efficiency by reducing the feature maps' sizes.
   iii) The fully-connected layer then receives results from the previous layer and sends it to the Neural Network layer.

c) Three examples of Convolutional Neural Networks:
   i) AlexNet
   ii) VGGNet
   iii) ResNet

3) What are the vanishing and exploding gradients problems in Backpropagation? Name several techniques to address these problems. (5 points)
    a) The vanishing gradient problem occurs as gradients can become smaller as they propagate backwards through the layers of a neural network. This means that the model will begin to learn very slowly. Some techniques for addressing these problems include:
        i) Weight Initialization
        ii) Batch Normalization
        iii) Activation Functions
    b) The exploding gradients problem is the opposite of the vanishing gradients problem. This means that the gradients become very large, leading to instability and divergence within the model. Some techniques for addressing these include:
        i) Gradient Clipping
        ii) Learning Rate Scheduling
        iii) Gradient Normalization

Question 2 (5 points):

Consider a learned hypothesis, *h*, for some Boolean concept. When *h* is tested on a set of 100 examples, it classifies 80 correctly. What is the 95% confidence interval for the true error rate for $Error_D(h)$?

1. Incorrect samples = total samples - correct samples = 100 - 80 = **20**
2. The percentage of incorrect samples = incorrect samples / total samples = 20/100 = **0.2**
3. The inverse for a 95% confidence interval = 1 - 0.95 = **0.05**
4. The z-score for a 95% confidence interval = NORMS.INV(0.025) = **1.96**
5. The margin of error = z-score x square root((percent incorrect x (1 - percent incorrect)) / 100) = 1.96 x square root((0.2 x (1 - 0.2)) / 100) = 1.96 x square root((0.2 x 0.8) / 100) = 1.96 x square root(0.16/100) = 1.96 x square root(0.0016) = 1.96 x 0.04 = **.0784**
6. 95% confidence interval for true error rate is 0.2 $\pm$ .0784 = [.1216, .2784]

Therefore, the 95% confidence interval for the true error rate for $Error_D(h)$ = **[.1216, .2784]**