Question 1.a

Assumption:

The dataset for each of the class for the MRI images can be considered as a Gaussian distribution mainly because of small variations in the class distribution. In addition, assuming that each image has a size of 256x256 pixels.

Prior Information:

In total there are 6 different classes including the CN tower images. The different classes of images are Head images, Neck images, Spine images, abdomen images, pelvis images and CN tower images in the dataset.

Approach 1:

1. Detecting the Outliers in Preprocessing step: considering the CN tower images as an outlier in the complete human body part images we can try to detect these outliers by using different outliers detecting techniques such as Z-score, IQR and Isolation Forest.

Approach 2:

If Approach 1 fails, we then use Soft Clustering method to cluster different classes in the dataset as Hard Clustering methods don't have uncertainity measures or a probability that tells us how much a data point is associated with a specific cluster. In our case we use Guassian Mixture Models (GMMs) for this problem. As the image dataset is a 256x256 we first flatten the matrix to a vector in the preprocessing step. In the second step we perform dimensionality reduction using Principle Component Analysis (PCA) to reduce the dimensions of the dataset. The main reason this step is required is due the compexity of GMM of $\mathcal{O}(NKD^3)$ as D = 256x256 the runtime complexity will be higher so to make it effecient we perform PCA by projecting the data on top principle components (can be choosen by considering the variance for the top Eigenvalues and Eigenvectors) to reduce the dimesion of the dataset and then pass PCA transformed dataset to the GMMs.

Gaussian Mixture Models (GMMs):

A Gaussian Mixture Model consists of several Gaussians which is given by $k \quad \epsilon \quad \{1, 2, .., K\}$, where K is the number of clusters in the dataset. From the prior information the total number of clusters is 6 so in our case K = 6.

For each of the Gaussian cluster k, the mixture has some important parameters as given below:

1. Mean (μ): The mean (μ) defines the center of the Cluster.
2. Covariance (Σ): The covariance defines the width of the distribution.
3. Mixing Probability (Π): It defines how big or small the Gaussian Function is.

Maximization algorithm can be used to obtain the optimal values of these parameters so that each Gaussian fits the data points belonging to each of the cluster. In general the Gaussian density function is given by

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{1}$$

In equation 1, x is the datapoint, D is the number of dimensions of each data point. Here μ is the mean and Σ is the covariance and N is the number of datapoints. As there are several Gaussians we need to find the optimal parameters for the whole mixture which can be modelled by considering that we want to know the probability that a given data point x_n comes form Gaussian k which can be expressed as below:

$$p(z_{nk} = 1|x_n) \tag{2}$$

This effectively tells us for a given data point x, what is the probability it came from Gaussian k. In equation 2 z is the latent variable which takes the value of 1 when x comes from Gaussian k else it takes 0 as the value. This variable is usefull in determining the Gaussian mixture parameters by calculating its probabiltiy of occurrence. Now let $\pi_k = p(z_k = 1)$ be the overall probability that a given point comes from Gaussian k and for k different Gaussian let $z = \{z_1, \ldots, z_K\}$. Now

$$p(x_n, z) = \Pi_{k=1}^{K} N(x_n|\mu_k, \Sigma_k)^{z_k} \tag{3}$$

solving this further after applying Bayes rule will yeild us

$$p(z_k = 1|x_n) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\Sigma_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)} = \gamma(z_{nk}) \tag{4}$$

using Expectation Maximization(EM) algorithm we can obatin the optimal parameters μ_k^* and Σ_k^* as below:

$$\mu_k^* = \frac{\Sigma_{n=1}^{N} \gamma(z_{nk}) x_n}{\Sigma_{n=1} N \gamma(z_{nk})} \tag{5}$$

$$\Sigma_k^* = \frac{\Sigma_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\Sigma_{n=1}^{N} \gamma(z_{nk})} \tag{6}$$

using EM algorithm we can converge to the likelihood value of the above two parameter which will effectively determine the cluster class for the data points and then we can identify the CN tower image from the rest of the MRI images.

Question 1.c

1. The computational complexity for the Guassian Mixture Model is $\mathcal{O}(NKD^3)$ where N is the total number of samples, K is the number of clusters and D is the dimension of the datapoint which will effectively reduce to $\mathcal{O}(N)$ for the case when $N >> K$ and $N >> D$. In addition, GMMs can be implemented to perform parallel computation so we can use GPUs to perform the computation as the size of the dataset is very large. This will reduce the computation time thereby giving results quickly.
2. We can use the above method for differentiating the MRI vs non-MRI images. If we have access to the labels for each of the dataset we can use different AI/ML models to train the network to identify the image otherwise we have to use unsupervised learning algorithms to differentiate between the two classes.