Method: KMeans, Louvian

Data: Pathology GAN, Inception V3

**Introduction:**

This case study is based on Cancer diagnosis by biopsy. This method includes removing a piece of tissue from the suspected region and analyzing it under a digital microscope. The data is on an average 20GB - 100000x100000 pixels. This data is passed through state of the art image classification techniques like PathologyGAN, Inception V3 to break the whole images to small patched and provide a label to the unlabeled data which is then passed through a dimensionality reduction algorithm like PCA, UMAP which reduces the complexity of the data by reducing its features.

The aim of this case study is to find out a best model, corresponding image processing and dimension reduction method which can classify the given labelled data and can be used to predict the presence on a tumor cells (colorectal adenocarcinoma epithelium)

**Methodology and Experimental framework:**

**Data:**

The data provided is processed through PathologyGAN and InceptionV3 classifiers respectively. The obtained vectors are passed through dimension reduction technique PCA: Principal Component Analysis and UMAP: Uniform Manifold Approximation and Projection. These methods brings the features down to 100. This make it easier to analyze and remove the features which has low or negligible impact on the classification.

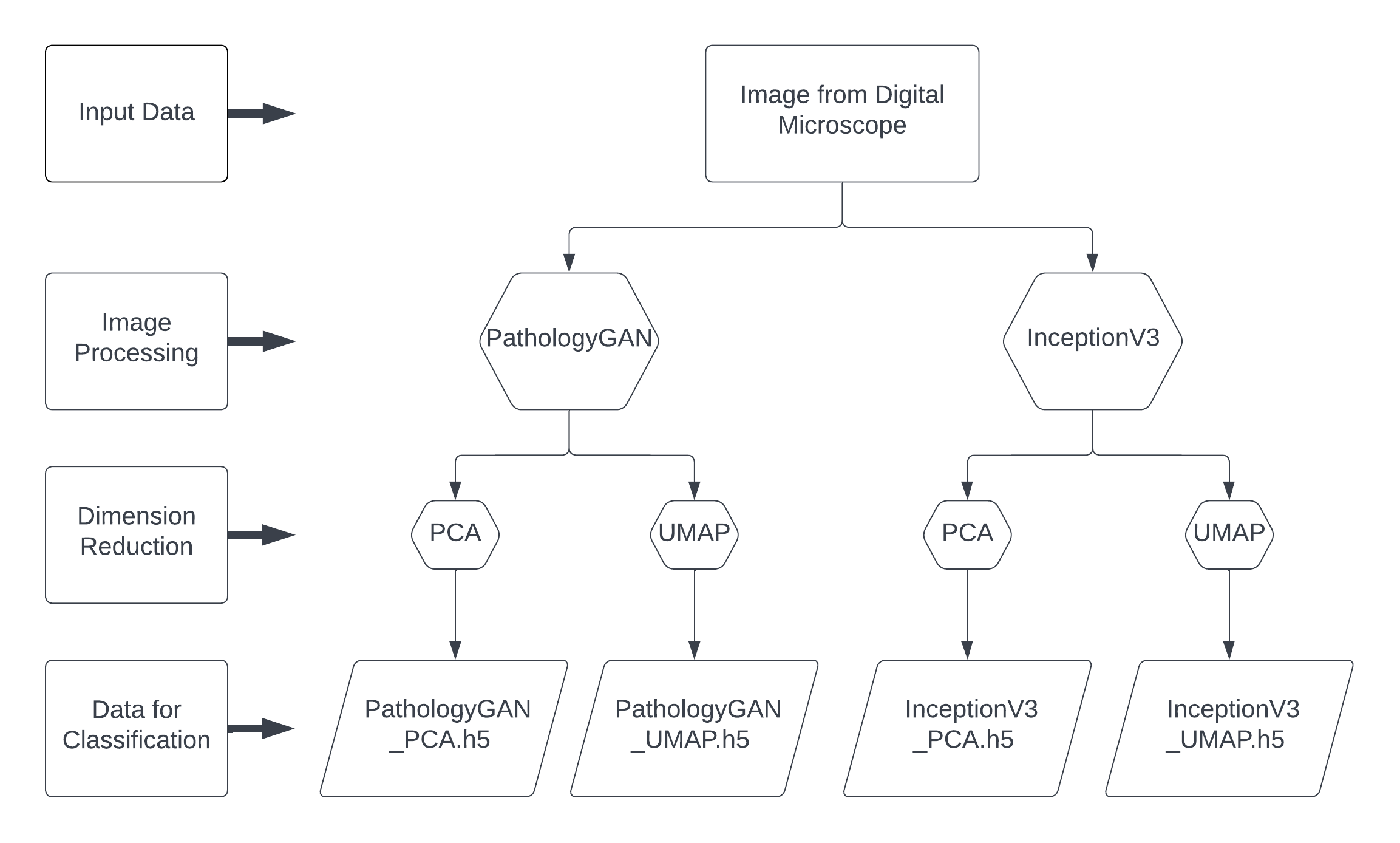


Figure 1: Data preparation

**Model:**

The clustering models used were K-Means and Louvain.

K-Means works on the principle of least squared distance (Euclidean distance), an initial guess of clusters has to be provided, which is placed at random location in the sample space and the distance between that location and the sample points are calculated, the newly assigned members of that group are used to calculate a new centroid and the process repeats itself unit convergence is reached, that is there is no change in the cluster membership**[1]**. Therefore the most important parameter is the initial guess of the number of clusters that we want the data to be sorted in. In this case study out initial guess will be from 2 to 30 clusters.

Whereas, Louvain clustering is primarily a graph-based clustering method, based on the principle of maximizing the modularity of clusters. In this technique each data point is represented as a node, and the similarity is represented by an edge. Various modularity formulas are used to assign nodes to a group the scikit-network has 3 modularity methods ‘Newman’, ‘Potts’ and ‘Dugue’ **[3]**. The resolution parameter determines the number of clusters formed. In this study we will be looking into Newman modularity method with the resolution of 0.9 to 1.1.

**Evaluation:**

The model evaluation method employed in this study is V-measure and Silhouette Score.

NEED MORE CONTENT

The V-measure use ground truth labels to measure the clusters obtained against Homogeneity (members of single class) and Completeness (members of same class are assigned to same cluster) **[2]**.

Silhouette Score is based on the distance of data point on a cluster with another cluster. This is calculated using the mean distance theorem. This score can be used to rank the performance of models, as the greater the score shows better clustering of the data points give.

**Result:**

The result with Louvian – inceptionv3\_pca detects tumor cells

Conclusion

Reference:

1. Rogers, S., & Girolami, M. (2016). A First Course in Machine Learning (2nd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315382159>

2. [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

3. [Scikit-network: Graph Analysis in Python](https://scikit-network.readthedocs.io/en/latest/reference/clustering.html?highlight=cluster), Thomas Bonald and Nathan de Lara and Quentin Lutz and Bertrand Charpentier, vol. 21 2020.