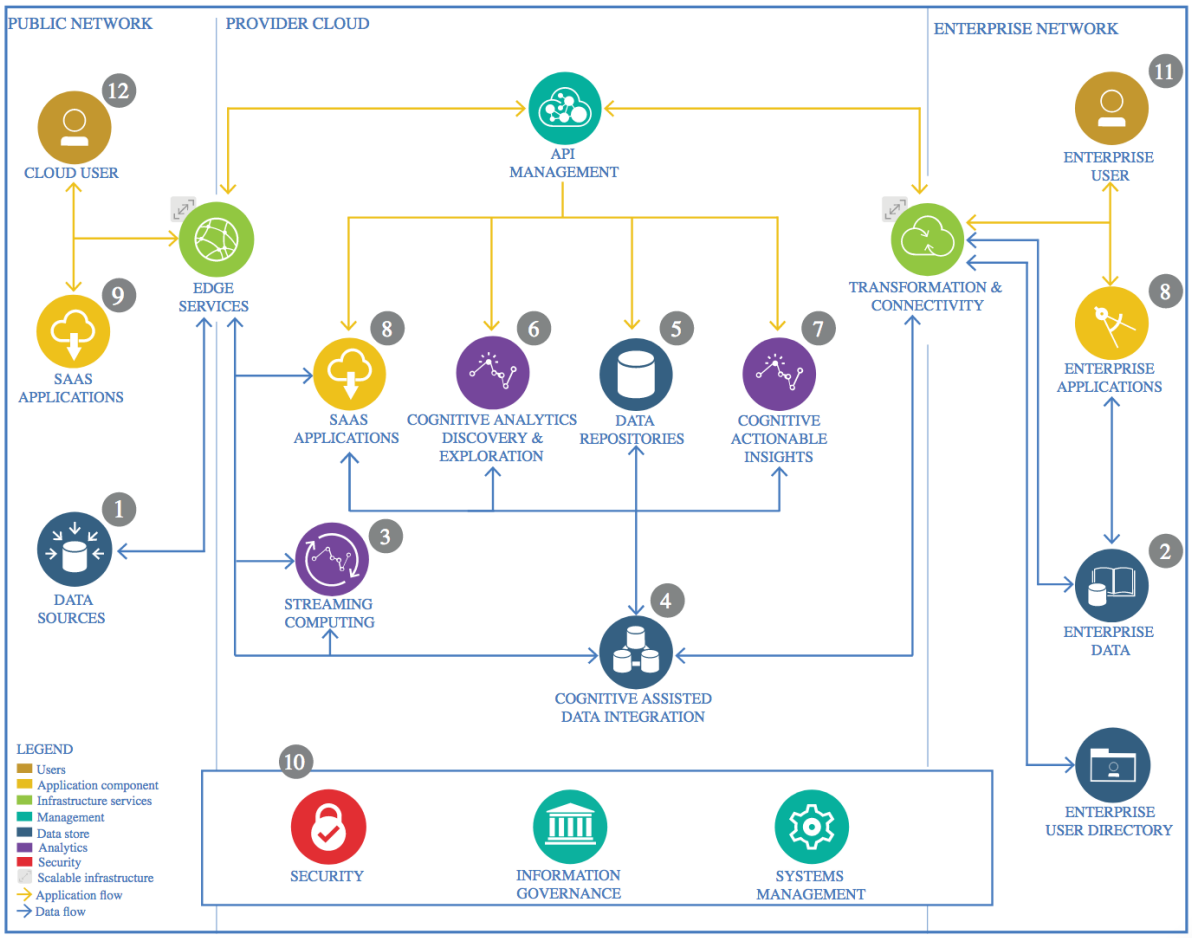
Utilizing Image Classification Techniques to Distinguish Cancerous Cell Tissue From Healthy Cell Tissue

Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

For the purpose of creating an algorithm that could predict whether images of lung and colon cell tissues indicated the presence or absence of a cancer, data was gathered from the works of Borkowski et al. (2019).

1.1.2 Justification

This data set demonstrates great consistency in the manner in which each cell tissue is measured, allowing for an ease of modeling and future prediction.

1.2 Enterprise Data

1.2.1 Technology Choice

The data utilized contains a directory of images organized by class. Each image was a 768x768 sized image, each containing 3 input channels. The data consists of 25,000 images with 5,000 each for following classes:

1) Benign Lung Tissue
2) Benign Colon Tissue
3) Lungs Squamous Cell Carcinoma
4) Colon Adenocarcinoma
5) Lung Adenocarcinoma

1.2.2 Justification

Given the high quality and resolution of these images, an image and its associated class appeared to be all that was required in order to create an efficacious algorithm for the classification of cell tissue as benign or cancerous.

1.3 Streaming analytics

1.3.1 Technology Choice

In this specific project, data streaming was not necessary. Instead, images were to be analyzed in batches. With this methodology the machine learning and deep learning procedures were trained using a set of images at a time. For the implementation of this procedure a data generator from the Keras function was utilized in order to read images from each image directory into each model iteratively.

1.3.2 Justification

The Keras library provided using the python programming language illustrated great efficacy in training models with a large amount of data that couldn’t ordinarily fit into memory. Using this method, the training of the model using batch processing did not require an amount of memory in excess of what was available.

1.4 Data Integration

1.4.1 Technology Choice

Given the large amount of samples in the data set used, it was not necessary to integrate outside data sources into this analysis.

1.4.2 Justification

Utilizing the 25,000 images available, it could be said with confidence that more data was not necessary to train a potentially effective model.

1.5 Data Repository

1.5.1 Technology Choice

Given the open source and HIPAA compliance of the data, the images used to train and validate the model were stored on a local file system in a directory denote “Images” for an ease of access. The data used to train the model are already hosted online at this location.

1.5.2 Justification

Given that the data size amounted to no more than 3 gigabytes, an Object Storage was not necessary to contain the data.

1.6 Discovery and Exploration

1.6.1 Technology Choice

For the purpose of performing this analysis, the frequencies of each class were first observed using a pie chart and a bar chart for data visualization. To explore the question of whether a model could be created for the purposes of predicting the class of each image, two models were devised. A traditional machine learning algorithm known as logistic regression as well as a deep learning algorithm known as a convolutional neural network as implemented.

1.6.2 Justification

Logistic regression is an established statistical and machine learning method that has proven useful in various areas of research. Similarly, convolutional neural networks exhibit a high degree of success in the areas of image classification. The efficacy of each model can be compared for the purpose of predicting the classification of each image.

1.7 Actionable Insights

1.7.1 Technology Choice

Subsequent to the feature engineering process in which each pixel intensity was divided by 255, each model was then evaluated in terms of its accuracy and F1 score. When comparing each model, it was determined that the convolutional neural network was most appropriate

for the task of classifying cell tissue as cancerous or benign from given each image.

1.7.2 Justification

Converting each rgb pixel intensity from an integer from 0 to 255 into a float ranging from 0 to 1 is a standard practice that results in an greater ease of mathematical evaluation. Evaluating each model, it could be observed that the convolutional neural network correctly classified the correct class of new data that it hasn’t been exposed to 92% of the time, while the accuracy of the multiple logistic regression model was accurate in its classification only 29.75% of the time. It could be said that the convolutional neural network was more appropriate for deployment.

1.8 Applications / Data Products

1.8.1 Technology Choice

Subsequent to the evaluation of each model, the convolutional neural network was saved into an *.h5 file for later use. A pair of helper functions were created for the purposes of later use on new histopathological images of lung or colon cell tissue.

1.8.2 Justification

As such model could potentially aid in the classification of future histopathological images, saving the resulting model into a file can result in an ease of later use and preclude the need to train the model before every instance in which it is to be used.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Given the open source nature of the project as well as the implausability of using the images to directly identify patients, security measures were not necessary to implement. In future circumstances in which the model could prove useful, it may be necessary for future practitioners to use data encryption methods or de-identification methods prior to its use when determining whether new histopathological images indicate th presence of a cancer or benign cell tissue.

1.9.2 Justification

HIPAA compliance within the medical field is a must for the protection of the identities of patients receiving care. Confidentiality must be protected as a result. While patients cannot be identified on the basis of these images or this model alone, future use on new data may require measures to be taken for the protection of patient identities.