The Potential of Artificial Intelligence to Identify Cancer and Aid Research

Author(s) Wilson B. Peguero Rosario

Master of Science in Data Science Capstone Design Specification

Grand Canyon University

Instructor: Professor TBA

Revision: 1

Date: December 7, 2022

**TABLE OF CONTENTS**

components testing 3

Requirements testing 5

system testing 8

**Components Testing**

|  |
| --- |
| ***Test Case Name:*** test\_pipeline |
| ***Priority:*** High |
| ***Module:*** pipeline.py |
| ***Test Objective:*** Test the functionality of the pipeline data extraction functions and assure that the output of the functions matches the input of the functions and algorithms on the models module. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step |  | Test Detail | Expected Results | Problem/Issue |
| 1 |  | The test\_extract\_data\_output function examines whether the output of the extract\_data function is a dictionary. | Datatype: dictionary | Dictionary datatype is required for model training and input. |
| 2 |  | The test\_transform\_data function examines whether the output of the transform\_data function has properly transformed the categorical data to its numerical representation. | The categorical values within the output are numerical in nature | Machine learning model observes categorical data as numbers (these are transformed into vectors of length n where n is the number of categories). |
| 3 |  | The test\_rescale\_image function examines whether the rescale\_image function rescales the image input by a factor of exactly two. Any larger reduction may cause a significant loss of data a lower factor may not be enough for data training. | Factor of 2 | The reduction in scale of an image must be exactly a factor of 2 |

|  |
| --- |
| ***Test Case Name:*** test\_model |
| ***Priority:*** High |
| ***Module:*** models.py |
| ***Test Objective:*** Examine the efficacy of the models as well as its ability to provide proper output. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step |  | Test Detail | Expected Results | Problem/Issue |
| 1 |  | The test\_model\_compilation function observes whether machine learning model can be loaded and set up for training. This function only examines whether the model has been successfully compiled or not | Model can make predictions post compilation | The model cannot compile for training. |
| 2 |  | The create\_test\_data function works to create a data set for testing different aspects of the model. | Test data set is created | Require data for testing the output aspect of the machine learning model. |
| 3 |  | The test\_model\_output function uses the create\_test\_data function to examine whether the tumor\_classifier function can create a model that outputs a list of probabilities based on classifications. | The output is a list of probabilities | The output of the model must be a list of probabilities to observe the model’s confidence. |
| 4 |  | The test\_model\_score\_probability function works to assure that the sum of the list of probabilities within the output equate to 1.00 | Sum of probabilities = 1 | The output is a list of probabilities that show the probability that the image contains a malignant or benign tumor. |
| 5 |  | The test\_model\_accuracy function iterates through a set of examples, creates a confusion matrix and calculates the accuracy using the pipeline module’s calculate\_confusion\_matrix function. It then asserts that the model accuracy is higher than or equal to 0.90. | Accuracy of model higher than or equal to 0.90 | The accuracy of the model must be at the very least 90% to compete against radiologists. |

**Requirements Testing**

There are a total of three general components, two of which will require testing. The three components are the data pipeline, the models module and the dashboard. Out of the three, only the data pipeline module and the models module require testing. Although the dashboard is provided and created, the main purpose of the algorithm is to develop a machine learning model that will classify image data, providing an api that allows one to use the machine learning model to make the predictions and display them on any dashboard app, such as python’s Dash library, or even Microsoft’s powerBI or Tableau will allow any institution interested to create their own machine learning model while easily displaying the results through the use of the data pipeline module.

|  |
| --- |
| Component:  Ability to Extract, Transform, and Load data for training |
| Name of Developer:  Wilson B. Peguero Rosario |
| Name of Reviewer:  TBD |
| Checklist   |  |  |  | | --- | --- | --- | | Type | Pass | Comments | | Functionality | **Passed** | Pipeline module is able to extract the data into a dictionary whose data is then transformed and loaded into a dataset object for model training. | | Versatility | **Passed** | The three functions to project ETL unto the data are interchangeable (meaning that they can be replaced independently of each other) while allowing one to save the training data for loading at another time. | | Adaptability | **Passed** | Able to select whether the model the data has been transformed for the first time or not to allow faster loading of data. | |

|  |
| --- |
| Component:  Ability to create machine learning model |
| Name of Developer:  Wilson B. Peguero Rosario |
| Name of Reviewer:  TBD |
| Checklist   |  |  |  | | --- | --- | --- | | Type | Pass | Comments | | Functionality | **Passed** | Model architecture can be used for any image or type of cancer. | | Performance | **Failed** | The machine learning model’s accuracy when viewing test data has increased from 63% to 71% by changing the from\_logits parameter to True. This still fails the 90% accuracy threshold. | |

**System Testing**

Text

Description automatically generated

**Figure 1:** Results of testing using the pytest framework.

Based on the testing set up for the model and the pipeline modules, the majority of the algorithm is ready for production (with the exception of the model accuracy). The ETL aspect of the data pipeline was tested together with the model compilation, training, input, and output. As mentioned before, the purpose of the dashboard is for standard display, but the purpose of the pipeline and models API is to create a machine learning model and feed the data to any dashboard application or library.