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: 3

Date: November 12, 2022

**ABSTRACT**

**[Special Note:** Instructions are provided in brackets. Delete these instructions prior to submission. For additional information and details, refer to the appropriate Capstone Project Handbook, located on the College of Science, Engineering and Technology page in the Student Success Center.]

[The abstract should summarize the entire project in 2 to 3 paragraphs, about 15–20 lines. The abstract addresses what the project is about, the tasks involved, and what will be accomplished. **Note:** Once the entire project is complete, this section will need to be updated just like the Project Proposal to provide a broader context of the major sections of the project (design, development, implementation, testing and overall functionality).]

|  |
| --- |
| **HISTORY AND SIGNOFF SHEET** |

**Change Record**

|  |  |  |
| --- | --- | --- |
| **Date** | **Author** | **Revision Notes** |
| August 8th, 2022 | Wilson B. Peguero Rosario | Initial draft for review/discussion |
| August 31st, 2022 | Wilson B. Peguero Rosario | Finished Draft for Submission |
| November 12, 2022 | Wilson B. Peguero Rosario | * Finished Design Planning summary by setting back the scope to only diagnose tumors based on malignancy. * Update description to include One Hot encoding of labels rather than ordinal encoding * Changed GPU hardware from 2060 to 3090TI |

|  |
| --- |
| **Overall Instructor Feedback/Comments** |

|  |
| --- |
| **Overall Instructor Feedback/Comments** |

**Integrated Instructor Feedback into Project Documentation**

Yes  No

**Project Approval**

*<Insert name of instructor here>*

**TABLE OF CONTENTS**

Design Planning summary 4

Overview of model pipeline design 5

Detailed model pipeline design 6

Appendix A – Technical Issue and Risk Log 9

Appendix B – References 10

Appendix C – Copyright Compliance 11

Appendix D – External Resources 12

Design Planning Summary

Cancer is an affliction that has existed since ancient times. Multiple solutions have been created to mitigate the harm that cancer does on the human body (such as surgery, chemotherapy, gene therapy, etc.). Although these solutions have assisted in mitigating cancer deaths, cancer remains as the second leading cause of death in the United States. Majority of the deaths are mainly caused by the late diagnosis in cancer (the point in which the tumors have spread across the entirety of the body). To assist in early diagnosis (while mitigate the yearly rate of misdiagnosis of 5%), machine learning algorithms can be employed. With machine learning, doctors can better observe and diagnose tumors as either benign or malignant.

**Project Deliverables:**

* Determine whether tumor is malignant or benign
* Provide probability or likelihood of tumor being malignant vs benign

**Overview of Model Pipeline Design**

To develop the pipeline, the data will be extracted from Google’s BigQuery open data sets from the NIH’s data commons website. After the query is done, a table containing links to the image data set will be downloaded by using python and the google-auth library to log in and access the images saved on Google servers. The images will then be loaded using the pydicom library as the dataset will contain a series of DICOM files (medical image format that contains textual information within the metadata of the image; DICOM stands for Digital Imaging and Communications in Medicine). The pydicom library will then be used to extract the textual data and the image, then TensorFlow’s Data API will be used to load and process the data. The image data will be processed using the preprocessing layer within the model while another preprocessing layer from TensorFlow will be used to convert textual data to numerical data. In the case that the textual data represents a label, the label will be encoded using TensorFlow’s OneHotEncoding functionality. This will allow the model to be trained without being misled. To visualize the data and observe patterns, one may grab a random sample to plot side by side with the different labels (i.e. a matrix of images at different stages of cancer). To analyze the model and interpret the results from the analysis, at least four models will be used. A base model with a single convolutional layer to observe the potential of Convolutional layers to make predictions, a model with preprocessing layers that will in turn normalize the data before processing to observe the effects of the preprocessing layer on the last model, a layer that interprets the contribution of the textual data to the explanatory variable and finally a composite model which takes in both processed image data and textual data to make predictions on the stage of cancer or whether the tumor is malignant or benign. A comparison between the metrics of all four models will be used and the predictions done will in turn highlight how the composite model may have its advantages over the independent models at making predictions or may suffer from the disadvantages of the worst performing model.

Detailed Model Pipeline Design

**Overview**

Cancer is the second leading cause of death within the United States. Although many cancers, such as cervical cancer, are completely treatable, many patients still die due to late diagnosis or even misdiagnosis. Although doctors have been able to identify and diagnose tumors based on measurements and medical images (i.e., Ultrasound, MRI, X-Rays, etc.), it can be difficult to determine with confidence whether a tumor is benign or malignant. The current rate of yearly diagnoses in cancer is 95% (1 in 20 people are misdiagnosed with cancer). Although it may be difficult for doctors to diagnose cancer on their own, machine learning algorithms can detect the most minuscule features related to the tumor in order to determine whether said tumor is malignant or benign. This may assist in decreasing the rate of misdiagnosis (which is currently 5% of patients diagnosed with cancer per year).

**The Data Sources**

The data sources will be the Cancer Imaging Archive (a set of public and private data sets containing hospital image data with annotations, diagnoses, etc.), The Imaging Data Commons (a more accessible version of the CIA which allows one to query and extract the data through Google’s BigQuery). These data sources are both backed by the NIH, meaning that the data has been curated and stems from reliable sources that use professional techniques to not only anonymize the data, but also to properly curate the data into a functional data set.

**Data types and Formatting**

The main data types are:

1. Image data
2. Textual Data

Both these data types will be loaded and extracted using pydicom and Google’s googleauth library from python to then process the images and the textual data for model training. The formatting of the data will be done through TensorFlow’s data api and machine learning layers. The images will undergo a processing layer called a convolutional layer for dimensionality reduction and feature extraction, while the textual data will be converted to numerical data.

**Data Cleaning**

Before extracting the features from the image, some images from the same sample may be removed and the data will be balanced as well. A separate data set containing image data with little to no description will be set apart from the data set used to train the data as one would expect to have data with similar textual annotations for each sample image provided.

**Initial Data Exploration and Visualization**

The initial data exploration will contain some metrics estimating the average word count of the annotations, the raw number of data points per label and a comparison of the images provided with the labels will be shown to demonstrate the difficulty of estimating the stage of cancer or whether a tumor is benign or malignant based on the image alone. The header metadata values selected to become part of the input for the model were selected based on the knowledge provided by studies in regard to factors leading to cancer. An example of this would be age (where for all genders the likelihood of having cancer increases as one ages), other factors include sex, weight, and more. Some other aspects related to the patient also impact likelihood that he or she may suffer from cancer, such as race, height, economic status and education (these factors are more blanket in nature as they encompass diet, mean height, exposure to radiation through microwave, etc.). All of these factors are evaluated based on the correlation matrix.

**Data Models and Nature**

There will be 4 models at the very least. The first three models will provide insight on the power of processing the image to make predictions, the impact of data augmentation and normalization, and the impact of feature extraction or dimensionality reduction. The first model will use the entire image to make a prediction, whereas the second model will use dimensionality reduction through convolutional layers to make a prediction, the third model will include only data augmentation and normalization and finally the last model will include a combination of the latter two models. The purpose of the last model is to not only save training time and model size, but also prevent the introduction of random noise through the image itself.

**Methodology**

The results will be interpreted based on the metrics provided by the TensorFlow library. Accuracy, loss in training, mean squared error, precision, recall, and other metrics will be used in tandem throughout the entire training procedure and testing procedure. A set of confusion matrices per model will be used to observe how the addition of certain factors may impact each model’s ability to recall and classify with precision. With the model complete, the user will upload the data through the Dash web app created. The image data and header data will then be extracted from the DICOM files uploaded to the app using two functions (one set to only extract the data from the DICOM file while another transforms the data and readies it for input). The application will then plot the data within six different charts portraying the data provided from the header files within the DICOM file together with the model’s prediction. At the bottom of the dashboard, is a table containing the metadata associated with uploaded DICOM file together with the prediction, and the probability of the tumor being benign or malignant.

**Configuration Changes**

To further develop the project some changes were in order. The first change to the pipeline is the manner in which data was fed to the system. As the machine learning model possesses two distinct inputs, data will need to be gathered within a dictionary before feeding into the machine learning model. As this is an online dashboard, not many great changes are required. The current API systems in relation to hospital EHR systems use the Fast Healthcare Interoperability Resources (FHIR) system that allows for the use of JSON outputs. This allows one to directly extract patient data and expand the categorical inputs of the machine learning model directly from the system. Meaning that any updates to the system or any historical validation of a patient may be used to further train the model and improve its accuracy. The change to this would be the manner in which one extracts the data, currently the data is downloaded as an extract from the database the related metadata is loaded from an excel or CSV file and finally transformed into a format that can be used for training and testing.

**Security**

As this application will be set locally to the environment, there will be no need for the application to meet any security standards. As a precaution the application will provide instructions on how to run the application locally similar other local plugins or applications (such as the OHIF viewer below):

[OHIF/Viewers: OHIF zero-footprint DICOM viewer and oncology specific Lesion Tracker, plus shared extension packages (github.com)](https://github.com/OHIF/Viewers)

The user or institution can activate a docker container image of the software and activate it into their local server for internal use only. Updates will be provided through GitHub or privately through the EHR system as a plugin.

**Hardware and Software Technologies**

The key hardware are the GPU, the memory, and the cpu, whereas the key software technologies are python, TensorFlow, pydicom, pandas, and numpy. Please observe the detailed explanation below:

1. The GPU (Asus TUF RTX 3090 TI)
   1. Although there are machine learning focused Graphics cards, the price range on the core performance is not worth buying for its functionality (the 3090TI provides great cooling solution while providing 24 GB for almost a third of the price of a machine learning focused RTX A6000 with 48GB of memory (The Costs for the NVIDIA RTX A6000 is $4,650.00 before taxes whereas the cost of the ASUS TUF 3090TI was $1613.31 after tax).
2. The CPU (AMD Ryzen 7 3800XT)
   1. The selection of the CPU is arbitrary as one used the GPU for the model training. The only component that would affect training in regard to the component affected would be the capability of the memory speed being locked due to the CPU being of a lower performance variety.
3. The Memory (Crucial Ballistix DDR4 4000MHZ)
   1. Required to transfer the model and data over to the GPU memory for training. The current size of the memory is 16 GB (two kits of 8GB) that may require expansion to 32GB to better match the size of the GPU memory.
   2. The speed of the memory was considered for allowing the user to comfortably transfer data without the memory speed being the bottleneck. Ideally, the best memory speed is somewhere between 3200-3600MHz with 4000MHz being able to sustain any heavy data processing tasks.

**Detailed Overview**

Graphical user interface, diagram, application

Description automatically generated

**Hardware and Software Technologies**

[Provide a detailed inventory of hardware and software technologies that will be used in the solution within the following tables. Make sure to:

1. List any frameworks or third-party libraries that will be used.
2. List any Proof of Concepts to be completed (POC) to ensure that the technologies and frameworks selected are the best fit for project’s purpose, cost, and solution. This section should also be updated with the purpose/rational for the POC and the results of the POC.]

|  |
| --- |
| **HARDWARE AND SOFTWARE TECHNOLOGIES** |
| 1.NVIDIA RTX 3090TI GPU |
| 2.AMD RYZEN 7 3800XT |
| 3.16GB of RAM |
| 4. python 3.10 or Julia |
| 5.Tensorflow or Flux |
| 6. pydicom library |
| 7. Dash library |

|  |  |  |
| --- | --- | --- |
| **PROOF OF CONCEPTS** | | |
| **Description** | **Rationale** | **Results** |
| 1. |  |  |
| 2. |  |  |
| 3. |  |  |
| 4. |  |  |
| 5. |  |  |

Appendix A – Technical Issue and Risk Log

[Identify and monitor project issues and risks.]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ISSUES AND RISK LOG** | | | | | | | | |
| **Issue or Risk** | **Description** | **Project Impact** | **Action Plan/Resolution** | **Owner** | **Importance** | **Date Entered** | **Date to Review** | **Date Resolved** |
| I/R | What is the issue or risk? | How will this impact scope, schedule, and cost? | How do you intend to deal with this issue? | Who manages this issue? |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

Appendix B – References

[List all project documentation references. List all references using APA style.]

Appendix C – Copyright Compliance

[For each external technical tool or code used, provide a reference to its copyright policy, clearly showing your right to use it. For each external technical tool or code used, detail how you used it, how you adapted it, how you modified it (if permitted), and why did you use it as opposed to writing your own. Only a small portion of your project may rely on external code. When code libraries/packages are used, explain why this was necessary/required/recommended. Seek instructor approval for using external resources prior to beginning to work on the project.]

Appendix D – External Resources

|  |  |
| --- | --- |
| **GIT URL:** | [wpeguero/capstone: Capstone Project information. This contains the algorithm used for the Master's in Data Science Thesis. (github.com)](https://github.com/wpeguero/capstone) |
| **Hosting URL:** | *(if applicable).* |
| **Screencast**  **URL:** | *(if applicable).* |