Computer Science Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of these areas so that you will have a complete and realistic overview of your project. Your course instructor cannot sign off on your project topic without this information.

Note: You must fill out and submit this form. Space beneath each number will expand as needed.

Any cost associated with developing the application will be the responsibility of the student.

INFORM INSTRUCTOR:

Potential use of proprietary company information: (Y/[✓ N])

ANALYSIS:

1. Project topic AND description:

   This project will take a dataset provided by the National Institutes of Health that contains approximately 110,000 chest x-ray (CXR) images taken with both posterior-to-anterior (PA) and anterior-to-posterior (AP) views as well as the associated findings of each x-ray by a trained radiologist. These findings consist of 15 different potential labels, 14 of which indicate cause for follow-up and 1 label indicating none of the 14 labels could be applied to the individual image. A machine learning model will be trained to analyze and classify similar images to determine the probability of diagnostic findings consistent with the given labels to assist in training new physicians or medical practitioners and to provide quick indications in critical care situations where immediate analysis of a film can mean the difference between life and death.

   The proposed client for this project is COMPANY, a fictional medical imaging company that provides on-call radiologic imaging readings for physicians in a variety of settings, ranging from urgent and emergency care to general practitioners in a small office setting. Chest x-rays are relatively inexpensive diagnostic scans compared to CT imaging, and do not require expensive or large machines in order to be performed. As such, COMPANY receives several CXR images on a daily basis for analysis and has had to devote a considerable amount of resources to providing results to their clients within a reasonable time frame. COMPANY is seeking a way to better streamline the process for providing a quick diagnostic finding for these images and lower the amount of human resources devoted to these scans so that they can be better applied to more complex and complicated imaging processes. By creating a machine learning model that can pre-screen x-ray images for the likelihood of a finding, COMPANY can implement a priority queue of scans and better devote time to those that contain diagnostic findings, confirm those without, and focus on images where the model is unable to determine, within a reasonable degree of certainty, that a classification can be made.

2. Project purpose/goals:

   The goal of the model will be to predict the presence of the aforementioned diagnostic markers in images to within a 90% likelihood of a correct classification. Each input scan will be output with a table of likely classification labels, the certainty to which these labels can be applied, and potentially highlight portions of the image that are indicative of the finding (i.e., by framing potential markers in a square, labelled box superimposed on the original image). The model will also be able to provide a highly accurate binary classification between FINDING and NOT FINDING to the end user, along with the probability that the classification is correct, so that those with a high degree of accuracy towards a FINDING can be further analyzed by both the model itself, and by trained radiologists.

   Initial supervised training of the model will be accomplished using the K-Nearest Neighbors (KNN) algorithm to classify each image to the provided diagnostic labels. The model will then be trained with a recurrent neural network to provide predictions and classification labelling on new input. The hyper-parameters of each model will be tuned and collected to provide additional classification information to the end user. This model will then be deployed as a web service API that will allow the user to upload a chest x-ray image and return the probabilities of each label classification.

DESIGN and DEVELOPMENT:

1. Computer science application type (select one):
   - Mobile (indicate Apple or Android)
   - [✓ Web]
   - Stand-Alone
2. Programming/development language(s) you will use:
   i. Python 3.8 for general purpose needs.
   ii. Python 3.8 – Anaconda distribution for machine learning models and data processing.
   iii. Python 3.8 with Flask web framework for web server functionality (*may not be required*).
   iv. Serverless Framework or AWS Serverless Application Model (SAM) for architecting and deploying back-end functionality.
   v. React.js framework (JavaScript/TypeScript) for presentation layer as a single-page application.
3. Operating System(s)/Platform(s) you will use:
   i. Ubuntu 20.04.1 (Linux OS) for prototyping and development.
   ii. Amazon Web Services (AWS) for hosting requirements.
   iii. AWS Simple Storage Service (S3) for static content storage and delivery (including front-end web application).
   iv. AWS Lambda for data processing and all back-end functionality and deployment.
4. Database Management System you will use:
   i. Amazon Aurora Serverless Database (PostgreSQL) for relational data (*may not be required)*.
   ii. Amazon DynamoDB (NoSQL document storage) for image metadata and non-relational data.
5. Estimated number of hours for the following:
   i. Planning and Design: 30-60
   ii. Development: 50-100
   iii. Documentation: 20-40
   iv. Total: 100-200
6. Projected completion date:
   December 10, 2020

IMPLEMENTATION and EVALUATION:

1. Describe how you will approach the execution of your project:
   i. Training data in the form of existing chest x-ray imaging will be collected along with the diagnostic determinations made by trained radiologists for each scan. The data will be collated for preparation to be used in machine learning models. Unusuable scans or data will be discarded and removed from training or test data.
   ii. A randomized selection of data (approximately 10,000 samples) will be used to train the models. A validation set of data will be set aside and used for evaluation.
   iii. An evaluation will be performed on the results of the training set. If a less than adequate modelling result is obtained, the model will be retrained with modified parameters or additional data until a reasonable accuracy has been obtained for classification.
   iv. Once a prediction model of 90% or greater is obtained, the model will be adapted to function as a web service for classifying new chest x-rays and providing diagnosticians additional information as scans are created in real time.
   v. Training data and prediction information will be collated for the purposes of data visualization regarding the model and the data used to train the model. This will be used as reference material to demonstrate the efficacy of the model in the presentation layer of the application.
   vi. A single-page web application will be created to display the data, data visualizations, and documentation created while developing the model and service. Additionally, a graphical user interface will be provided for interacting with the prediction model service that allows a user to provide their own sample data and allow the model to make predictions. Unused training data will be set aside for demonstration purposes to allow a user without data to interact with the service and get a full demonstration of how to use the service.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

STUDENT SIGNATURE

November 3, 2020

By signing and submitting this form, you acknowledge any cost associated with development and execution of the application will be your (the student) responsibility.

COURSE INSTRUCTOR'S NAME:

COURSE INSTRUCTOR APPROVAL DATE: 11/3/2020