UNIVERSITY OF ABERDEEN

University of Aberdeen

School of Natural and Computing Sciences

Department of Computing Science

MSc in Artificial Intelligence

2020 - 2021

Please read all the information below carefully
Assessment Item 1 of 2 Briefing Document – Individually Assessed (no teamwork)

CS551G - Data Mining and Visualisation	***Note: This assessment accounts for 50% of your total mark of the course.***

Learning Outcomes

On successful completion of this component a student will have demonstrated competence in the following areas:

- Understanding of goals relating to different types of data mining techniques, ability to identify appropriate goals for extracting information from different data sets, and ability to apply all this in practice.
- Understanding of key models that support data mining, and ability to use appropriate models in practice.
- Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple data mining, visualisation and machine learning strategies

Information for Plagiarism and Conduct: Your submitted report and source code may be submitted for plagiarism check (e.g., Turnitin). Please refer to the slides available at MyAberdeen for more information about avoiding plagiarism before you start working on the assessment. Please also read the following information provided by the university: https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/

In addition, please familiarise yourselves with the following document "code of practice on student discipline (Academic)": https://tinyurl.com/y92xgkq6.

Application Problem Definition: Generating Synthetic Images and Detecting Covid-19 from Chest X-Ray Images

In the early days of the Covid-19 pandemic, one of the diagnostic tools available to clinicians was chest X-ray images (CXRs). CXRs (figure 1) remain one of the main imaging tools to evaluate Chest diseases, including Covid-19, bacterial pneumonia, etc.

This assessment aims at building a classification tool for detecting Covid-19 and also generate synthetic images that can be used to augment the images available.

Please read all the information below carefully

The dataset that includes Covid-19 and Normal (no-disease) images can be downloaded from MyAberdeen. No prior knowledge of the domain problem is needed or assumed to fulfil the requirements of this assessment.

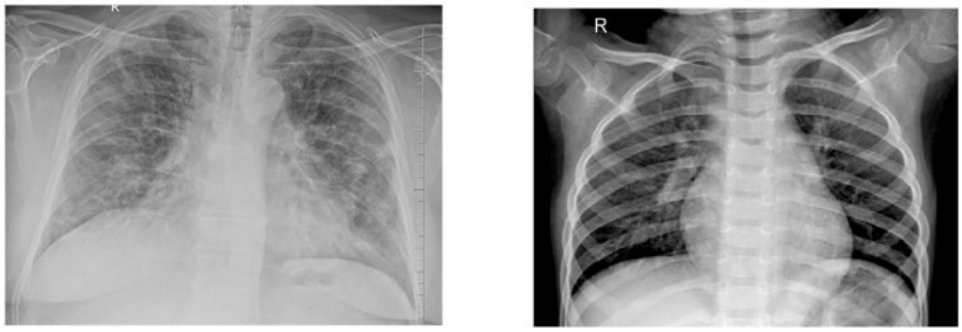

Figure 1. Covid-19 (left) vs Normal (right) X-Ray Images

Report Guidance & Requirements

Your report must conform to the below structure and include the required content as outlined in each section. Each subtask has its own marks allocated. You must supply a written report, along with the corresponding source code written in **python (CoLab or Jupyter Notebook)**, containing all distinct sections/subtasks that provide a full critical and reflective account of the processes undertaken.

Task 1: Create Synthetic Covid-19 X-Ray Images with Conditional Generative Adversarial Networks (35/50) – (max ~1500 words)

For Task 1, you are provided with the cgan.py class, which you can adapt to the problem at hand. Therefore, you do not have to implement the cGAN architecture from scratch; but you would have to play with the hyperparameters, such as learning rate, number of epochs and batch size to optimize the outcome, i.e. generated image quality.

You are asked to do the following:

Subtask 1.1: Using the Covid-19 dataset provided (100 images) and the cgan.py class, generate 50 synthetic images. You need to create a script that preprocesses the data, trains the cGAN model, generates synthetic images and saves them to a folder for further exploration. For your reference examples of a synthetic images can be seen in figure 2 (**27 Marks**).

Figure 2. Synthetic Covid-19 Images generated with Generative Adversarial Networks.

Please read all the information below carefully

Subtask 1.2: Generate synthetic images based on three different hyperparameters, e.g. different learning rates, epochs, batch size, etc. Clearly show the generated images according to the hyperparameters chosen. For instance, you might show generated images at epoch 100 and also at epoch 200, and so on (**8 Marks**).

Task 2: Detect Covid-19 from Chest X-Ray Images using pre-trained networks (15/50) – (max ~ 800 words)

Using a pretrained ResNet-50 model that can be found in https://keras.io/api/applications/ and also reporting performance in terms of accuracy, do the following:

Subtask 2.1: Via a transfer learning process, use a pre-trained ResNet-50 model (with imagenet weights) to develop a binary classification model that can classify Covid-19 and Normal CXRs. To accomplish this, use the 100 Covid-19 and 100 Normal CXRs images provided to you. Follow a 80% (train) / 20% (test) process (**hint: freeze the whole pretrained model, remove the last fully connected layer and add two new trainable layers.) (10 Marks).**

Subtask 2.2: Repeat subtask 2.1, but now randomly select 50 of the real Covid-19 images and add another 50 images generated as part of Task 1. The normal images remain the same 100 ones used in subtask 2.1 **(5 Marks)**.

All results reported shall be based on the performance achieved with the test set.

Marking Criteria

- Quality of the report, including structure, clarity, and brevity
- Reproducibility. How easy is it for another MSc AI student to repeat your work based on your report and code?
- Quality of your experiments, including design and result presentation (use of figures and tables for better reporting)
- Configured to complete the task and the parameter tuning process (if needed)
- In-depth analysis of the results generated, including critical evaluation, insights into data, and significant conclusions
- Quality of the source code, including the documentation of the code

Submission Instructions

You should submit a PDF version of your report including code snippets via MyAberdeen by **23:59 Sunday 28th March 2021**. The name of the PDF file should have the form "CS551G_Assessment2_< your Surname>_<your first name>_<Your Student ID>". For instance, "CS551G_Assessment2_Smith_John_4568985.pdf", where 4568985 is your student ID.

In addition to the written report, you should also submit supplementary material in the form of a zip file containing the source code of your implementation (ideally as a python notebook ".ipynb"). The naming convention should follow the same form as for the PDF. For instance, "CS551G_Assessment2_Smith_John_4568985.zip", where 4568985 is your student ID.

Please read all the information below carefully
Please try to make your submission file less than 20MB as you may have issues when uploading large files to MyAberdeen. Any questions pertaining to any aspects of this assessment, please address them to the course coordinators Milan Markovic (milan.markovic@abdn.ac.uk) and Aiden Durrant (a.durrant.20@abdn.ac.uk).