COURSEWORK OVERVIEW

Module Number & Title	**CMM536 – Advanced Data Science**
Submission Method	*In the COURSEWORK PART 1 (FORMATIVE) DROPBOX on the module's CampusMoodle page, you should submit:* - *A .ipynb file or link to an online jupyter notebook (in a .txt file). All cells should be executed and output clearly shown.* *In the COURSEWORK FINAL DROPBOX on the module's CampusMoodle page, you should submit:* - *A .ipynb file or link to an online jupyter notebook (in a .txt file) All cells should be executed and output clearly shown.*
Deadline	Formative Submission: Friday 4th March 2022, 4 pm Final (Summative) Submission: Friday 22nd April 2022, 4 pm **The submission deadline is 4pm** – whilst a 30-minute grace period has been added for technical issues, you should aim to submit by 4pm.
Module Co-ordinator	*Carlos Moreno-García* *c.moreno-garcia@rgu.ac.uk* For Staff Office Hours, visit: http://campusmoodle.rgu.ac.uk/course/view.php?id=96625
LOs Assessed	This coursework assesses all LOs: 1. Critically appraise the challenges posed by the management and processing of complex datasets and data inputs. 2. Discuss, compare and contrast advanced techniques and algorithms for working with complex datasets and data types using data science. 3. Critically evaluate and select state-of-the-art data science techniques and algorithms for selected/given applications involving complex data. 4. Apply advanced techniques and algorithms and critically analyse and evaluate the results.

AUXILIARY INFORMATION

Please refer to the coursework brief below for detailed information on your submission, including any software packages/versions that should be used, word counts and limits.

Coursework received after the submission deadline indicated above will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost. Coursework extended due to extenuating circumstances shall be assessed in the normal way.

If the **word count** of an assessment is considered critical, then this will be reflected in the assessment criteria for that assessment together with any consequent penalties.

In line with the RGU Assessment Policy, **this coursework has been moderated** by the School of Computing Moderation Panel, which comprised a detailed technical review and panel overview to schedule staggered submissions.

ACADEMIC INTEGRITY

All work is expected to be completed by yourself, unless you have been clearly instructed to work as part of a team or group. Before submitting assignments, you should check your submission to ensure that it complies with Academic Regulation A3-2 Student Conduct Procedure and Academic Integrity, including but not limited to:

- *all material identified as originally from a previously published source has been properly attributed by the inclusion of an appropriate citation at the point of use in the text;*
- *direct quotations are marked as such (using "quotation marks" at the beginning and end of the selected text), and*
- *full details of the reference citations have been included in the list of references.*

EXTENSIONS AND DEFERRALS

The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so. For information about extensions and deferrals, please familiarise yourself with the extenuating circumstances defined therein.

Please read the entire coursework specification carefully before starting the coursework. If any aspect of what you are being asked to do is not clear, seek advice and assistance from the Module Coordinator.

1. Background

Nowadays, medical image analysis is one of the most recurrent uses of modern data science. Thanks to the recent advances in computer vision and machine learning, millions of images can be collected and processed in a matter of seconds to determine the presence or absence of a certain disease. As a result, you will find a vast number of scientific papers and projects which tackle issues such as fracture detection [1], cardiac segmentation [2], COVID identification in chest x-ray images [3], amongst others.

2. Focus Area

In this project, you will design a comparative experimental validation framework to test different computer vision and machine learning algorithms which will help you **classify x-ray medical images from different body sections.** Your coursework consists of:

Part 1 - Data collection: To begin, you need to compile your own dataset based on AT LEAST THREE x-ray image repositories available online. Remember, the purpose of this coursework is **not** to classify the disease or medical condition, but rather to create a system that is capable of distinguishing between different types of body parts! For instance, you may find a data repository which was used to distinguish between fracture and no fracture in wrist radiographies. For your coursework, you would need to collect *all of these images (both positive and negative cases)* and label them as "wrist". Afterwards, you can look for other images, such as chest, arm, head, or other x-rays, and label them accordingly.

One of the main requirements of your collected dataset is that **it must be imbalanced!** For example, if you have collected 1000 wrist images, you are encouraged to look for other x-ray repositories which contain more/less images. As guidance, to obtain acceptable results in the validation stage without demanding too much computational power, I recommend you to have at least 500 images per class.

The collection of your dataset can start as a "manual process" (i.e. downloading images and storing them in folders in your computer). However, you need to show in your notebook that you can import these images into Python and create a numpy array/pandas data frame where each row contains each **grayscale** image and the class/label of such image. Also, don't forget to reference the original data repositories from where you obtained the images to be used.

I recommend you to be careful with the format of the images collected. In the medical image domain, you will find that many people work with formats such as DICOM which are very specific and require specialised software to be manipulated. Instead, I would favour the use of "classical" image formats such as jpeg, png, etc. Still, you will see that with a simple google search you will find plenty of publicly available options!

Part 2 – Validation: Once that you have created the data structure of your "starting" data repository, apply a stratified data split of the data so that 70% of your dataset can be used for training and 30% for testing. Afterwards, you need to carry out the following four experiments obtaining the precision, recall and f1-score in each one.

a. Training with the training data unaltered and classification of the test set using a neural network-based machine learning architecture.
b. Balancing the classes of the training data (e.g. using image augmentation) and classification of the test set using a neural network-based machine learning architecture.
c. Feature extraction of the starting dataset (e.g. Harris, SIFT, SURF, HOG, etc.), and classification of the test set using any machine learning architecture (e.g. NN, CNN, SVM, Random Forests, Naïve Bayes, etc.)
d. Feature extraction of the starting dataset (preferably the same used in c.), data balancing of the training dataset (e.g. oversampling, undersampling, class decomposition, etc.) and classification using any machine learning architecture (e.g. NN, CNN, SVM, Random Forests, Naïve Bayes, etc.)

Afterwards, create a table or plots where all of the results are compared. Discuss which method/combination of methods is better, if you noticed a trade-off in the performance, and your thoughts on how the work could be improved provided that you had more time, computational resources, etc. You can also rely on other metrics or outcomes, such as the confusion matrix, ROC AUC, etc.

Finally, the report structure and presentation will be considered for the final mark. You must produce a clear and practical notebook that is easy to navigate (i.e. do not print unnecessary or long outputs) and that shows all code used with markdown cells which comment and discuss the findings of each step. The total word count of the markdown cells must not exceed 1500 words.

3. Deliverables

Formative feedback will be provided for part 1 (see the cover for hand-in dates).

Afterwards, you need to present a single report where the whole work is presented.

4. How to calculate your final mark?

Each of the following aspects of the final report convey to a different number of marks:

- Part 1 (Data compilation): 2 marks
- Part 2 (Validation - experiments): 3 marks
- Part 2 (Validation – discussion of results): 2 marks
- Report structure and presentation: 1 mark

Your mark will be calculated from the following grading profile:

Grade A

At least 4 subgrades of A.

At least 6 subgrades of B or above.

All 8 subgrades at C or above.

Grade B

At least 4 subgrades of B or above.

At least 6 subgrades of C or above.

All 8 subgrades at D or above.

Grade C

At least 4 subgrades of C or above.

At least 6 subgrades of D or above.

Grade D

At least 4 subgrades of D or above.

At least 6 subgrades of E or above.

Grade E

At least 6 subgrades of E or above.

Grade F

The report has been submitted, but the set of subgrades does not qualify for any higher grade.

The marking grid for each of the coursework elements can be found in Moodle.

5. References

1. Moreno-García CF, Dang T, Martin K, Patel M, Thompson A, Leishman L, et al. Assessing the clinicians' pathway to embed artificial intelligence for assisted diagnostics of fracture detection. In: Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data, co-located with 24th European Conference on Artificial Intelligence (ECAI 2020) [Internet]. Santiago de Compostela; 2020. p. 63–70. Available from: http://ceur-ws.org/Vol-2675/paper10.pdf

2. Dang T, Nguyen TT, McCall J, Elyan E, Moreno-García CF. Two layer Ensemble of Deep Learning Models for Medical Image Segmentation. ArXiv [Internet]. 2021; Available from: http://arxiv.org/abs/2104.04809

3. Nikolaou V, Massaro S, Fakhimi M, Stergioulas L, Garn W. COVID-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network. Health Inf Sci Syst. 2021 Oct 12;9(1):36. doi: 10.1007/s13755-021-00166-4. PMID: 34659742; PMCID: PMC8509906.