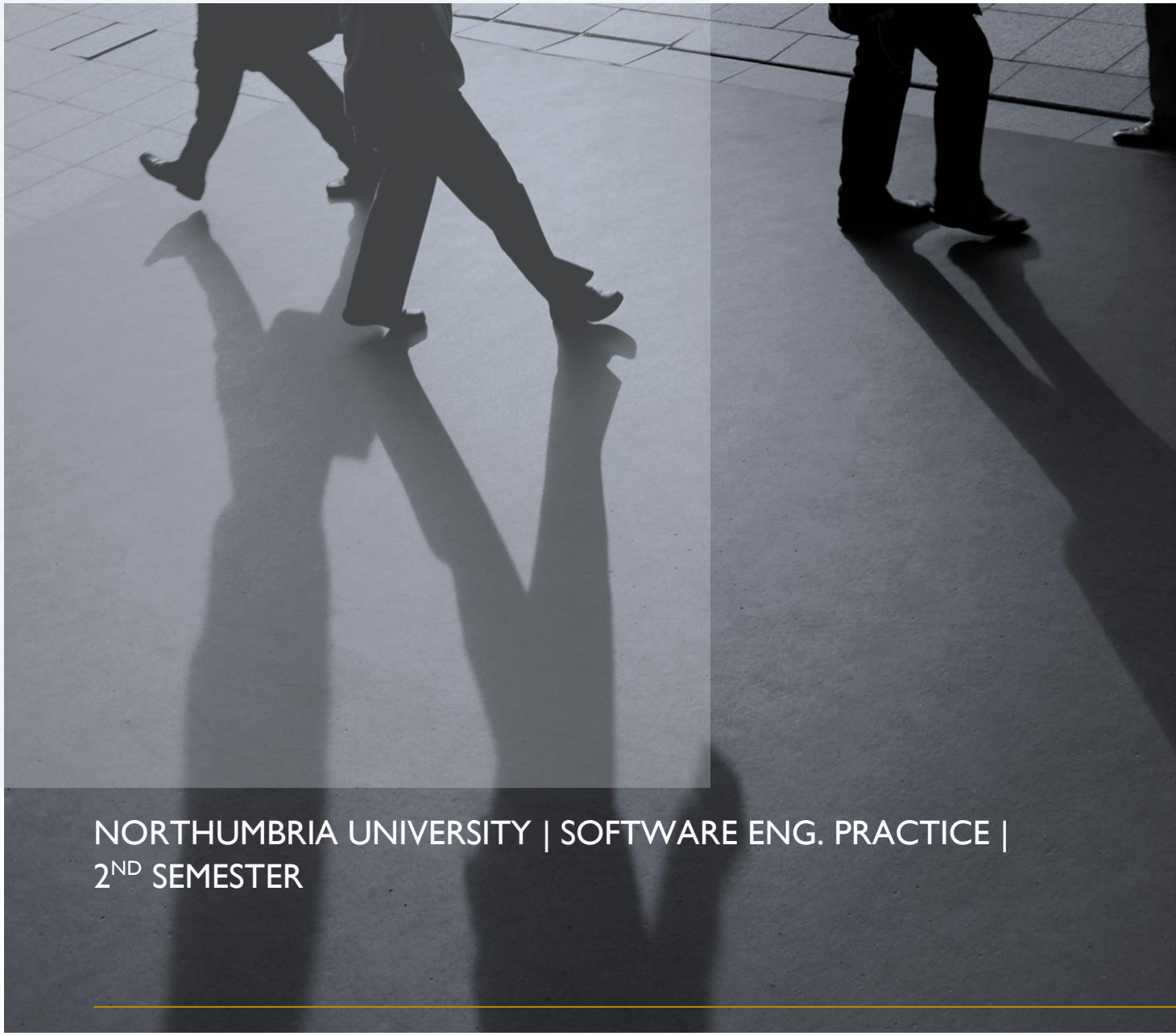
TEAM HORSE CHESTNUT

ADDITIONAL DATA



NORTHUMBRIA UNIVERSITY | SOFTWARE ENG. PRACTICE | 2ND SEMESTER

TABLE OF CONTENTS

INTRODUCTION

This document refers to the Additional Data mission by team Horse Chestnut of the AI stream of the Software Engineering Practice module at Northumbria University. With Iterative Development done, more data was to be added to increase the useability and robustness of the model. As such, the OASIS dataset was chosen, as it has the same amount of well specified categories. While this dataset is not particularly similar to the original, it still refers to the same condition and exhibits some similarities.

ADDITIONAL DATA LOADING AND PREPROCESSING

Since the new images were not like the ones used during Iterative Development, some pre-processing was to be done. When loading the data, only 488 images were used for each class, with this being the number of images of the smallest class. This was done so as not to accentuate the differences in class distribution of the original dataset, considering that this new data has around 60000 observations. After importing those images, they were rotated 90º to the left for them to follow the same orientation of the original images. They were then split into train and test sets and concatenated to those of the original images, being further pre-processed with the rest of the images.

JUSTIFICATION

While the considered models showed promising results with the pre-processed MRI scans, those can be hard to find. The OASIS dataset is the largest collection of MRI scans, including multiple layers of the brain of several patients. This dataset, contrary to the original one, has features such as the eyes and the skull still included. While their removal could be beneficial, it was decided that they would be kept, to make the model more robust and able to work with more types of scans. This means that the model has images both with and without these features and is thus not as limited in terms of the data it can work with.

As such, as was one of the established objectives of the team for this mission, the idea was to not only add more images to the dataset but to also attempt to give more data for the model to learn about the minority class, but to make it more generalisable, as said before, and not as limited as it was before. The advantage of this methodology is that new images that are given to the model do not have to follow the same pre-processing pipeline as the originals, as the model can now deal with a more differentiated sample. Although it is to be expected that performance might suffer, it is still important to consider a less restrictive model.

As every class is receiving the same number of images, the model has more data to learn from without making the distribution more unbalanced. While the images do not contain the exact

same data nor features, it is still expectable that similarities exist, as they both come from MRI scans from the same angle.

The choice of the OASIS dataset over another, such as ADNI, was due to the fact that the former also has 4 classes with similar names to the original. The latter also has 5 classes that do not have a clear match in the original dataset, which made it not worth using as there is no way to guarantee that the data is consistent. What appeared to be a heavily augmented version of the original dataset was also briefly considered, but as the augmentation was carried out without splitting the dataset, there was bound to be data leakage between the sets, which rendered it unusable. All these datasets were found using Kaggle.

RESULTS AND DISCUSSION

[Figure 1 – Table of Accuracy and Sample Size between the models]

Type of Model	**Name of Model**	**Accuracy of Alzheimer's Classification (%)**	**Accuracy of Multiple Classification (%)**	**Sample Size**
Pretrained CNN (MRI)	**XbADM_Large (Ours)**	80.3	70.7	7552
Pretrained CNN (MRI)	**XbADM_Base (Ours)**	83.9	91.07	5600
CNN (MRI)	**CbADM_Large (Ours)**	89.7	80.1	7552
CNN (MRI)	**CbADM_Base (Ours)**	98.2	94.07	5600

After running the code with both balanced and unbalanced dataset, the results that were given by the model were worse than those of the original, which was expected, as the new dataset has been proven to be considerably harder to predict and has different features to the original dataset. Even still, the models still performed well and got around 80% with the balanced data and all the 4 classes on 25 epochs, though the self-made CNN (*CbADM*, which stands for CNN based Alzheimer Detection Model) seemed to suffer more than the pretrained (*XbADM*, which stands for Xception based Alzheimer Detection Model) model on binary data, with the opposite being true for multiple classification. Changing the number of epochs, the optimiser and adding more layers could help the models improve, but for the sake of consistency they were kept the same as the ones from Iterative Development. Even with a set seed, the performance of the models still varied each run, so the difference can be more or even less drastic than the one present. It is important to note that the pretrained model was still learning more about the data, so it could be expected that it would possibly surpass the *CbADM*, which would achieve that accuracy early on

before plateauing. As such, it is important to decide between a faster model or a slower one that could be more precise if given more time to learn.