INTRODUCTION

The red-dot model being evaluated consists in inputting Chest X-Rays images and scanning through the different images to identify any biological abnormalities. Images are then further categorised into two classes: Abnormal and Normal. The main challenge in automated medical diagnosis is to ensure all patients have the same likelihood to be diagnosed at the earliest stage possible and therefore treated equally. This prompted us to ensure the good performance does not vary based two major determinants: patients’ genders and manufacturers. The manufacturers depict 11 different brands capturing the images. The gender counts for females, males and non-binary patients. The first step to this investigation is to explore three given datasets. The first set of data comprises an ID number unique to the patient’s medical exam, an accession number (AN) proper to a specific gender and manufacturer, and the corresponding gender and manufacturer types. The second set of data provides radiologists’ diagnosis which categorises the AN exam as abnormal or normal. The third set of data is what the red-dot model outputs when scanning through the images and identifying potential abnormalities depicted as a red dot on the image. Combining all three datasets in the most accurate way will allow us to define the model’s performance accordingly to the different determinants mentioned above.

\*red-dot model

METHOD

As the determinants were not provided for all datasets, in the case AN were duplicated, it was impossible to identify their respective gender and manufacturer types. Therefore, removing duplicates was the first step to data cleaning. To avoid evaluating ANs where not all information were provided, we removed ANs not present in all three datasets. Combining all datasets together make information available for each AN remaining. Counting the number of unique gender and manufacturer type resulted in the understanding that some determinants were poorly present in the data. For instance, only one non-binary has been recorded and only 1 image was generated by the Varian manufacturer. Those were removed from the study as it could affect the analysis when the model performance would only be investigated on only a few samples. The initial 15 manufacturers were grouped together where these were comprised as similar corporations yet preventing small samples evaluation. Finally, all ANs remaining were inputted in the statistical analysis. The model’s performance was defined as good when both the algorithm output and the radiologist’s diagnosis were in agreement. Inversely, if the algorithm categorises the image differently than the radiologist’s saying, the model was considered as badly performing. The relationship between individual determinant and the model’s performance was then investigated by conducting a logistic regression, as statistical evidence.

RESULTS

Graphical illustrations were displayed to help visualising the model’s performance related to the image’s characteristics. Figure 1 represents the number of images classified as Abnormal (i.e. False) and Normal (i.e. True) by the machine learning algorithm for distinct patient’s gender. Figure 2 represents the number of images classified as ‘Abnormal’ (i.e. False) and ‘Normal’ (i.e. True) by the machine learning algorithm for distinct manufacturers. Figure 3 represents the statistical predictions for which the model performs more or less badly in categorising the image according to the patient’s genders and manufacturers. Overall, the red-dot algorithm tends to categorise the image of a male abnormal more frequently than female, and inversely, as normal for more female than male. Additionally, FUJIFILM Corporation, Philips Medical Systems, Canon Inc. and Samsung Electronics have a noticeable higher proportion in one of the two categories. The rest of the manufacturers are balanced between both classes. Finally, from the regression’s output we can clearly state that the model’s performance increases when analysing images from Samsung Electronics and Carestream Health, as well as for female than male. Manufacturers on which the model performs badly are depicted as Canon Inc. and Philips Medical Systems, as well as male. However, its performance on the rest of the manufacturers do not appear to have a strong relationship and seem to be quite balanced between poor and good performance. These results could suggest biological variations between female and male are dependent factors for the model’s output, therefore giving more chances for women to have better diagnosis than male. Another concluding result is that certain manufacturers quality of images tend to affect the model’s performance, therefore leading the model to vary its accuracy accordingly.

Chart, bar chart

Description automatically generated

*Figure 1: Number of abnormal and normal diagnosis according to the patient’s gender*

Chart, bar chart

Description automatically generated

*Figure 2: Number of abnormal and normal diagnosis according to the image’s manufacturer*

Chart

Description automatically generated

*Figure 3: Model’s performance for various determinants*