HeartLab AI Test Report

Executive Summary

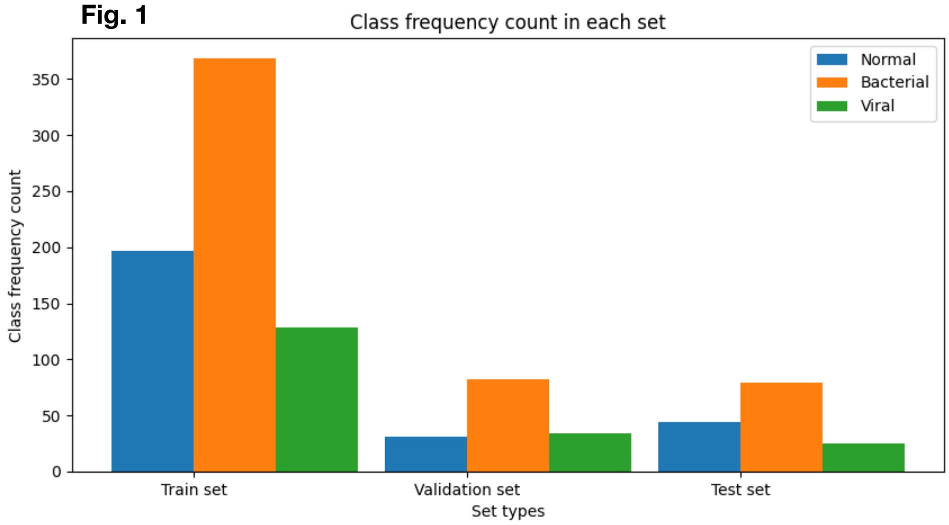
We have developed a CNN model to classify pneumonia X-ray images into three classes: normal, bacterial or viral. Our model has achieved 80% overall accuracy with macro-average precision and recall of 78% and 75%, respectively.

We have conducted data preprocessing unit tests and model resiliency tests, which showed that the model is resilient to image noise and rotation, achieving 74% accuracy on a noisy dataset.

Data Preprocessing

In the first step of data preprocessing, we have split the dataset into training, validation and test sets. We have used 70% of the data for the training set, 15% for the validation set and the remaining 15% for the test set. Before the split, the data was randomly shuffled to ensure more even class distribution in each set. The plot [Fig. 1] shows that class distribution is similar for all sets, with bacterial pneumonia being the largest class in every set.

Fig. 1



One-hot-encoding was used for dataset labels to allow the training of a CNN model.
Prior to training the model, we have normalised the dataset values between 0 and 1 to reduce the chances of exploding activation function coefficients and dying neurons.

Model Structure

We have used a CNN model for this image classification task. The model consisted of three layers consisting of convolutional, max pooling and dropout layers. After the three layers, the images were flattened, and a dense layer was used. For each of the above layers, RELU activation function was used.

The final layer of the model was a dense layer of three neurons with SoftMax activation function, where each neuron represented one of the three pneumonia classes.
Categorical cross-entropy and Adam optimiser were used to train the model.
We have used early stopping when training this model to reduce the chances of overfitting.
We have not implemented batch normalisation for the model since we have assumed that it is not deep enough for batch normalisation to provide a significant improvement.

Model Performance

The training accuracy [Fig.2] and loss [Fig.3] plots show no evidence of overfitting since there is no significant disparity between validation and training accuracy, and there is no increase in validation loss.

Fig. 2

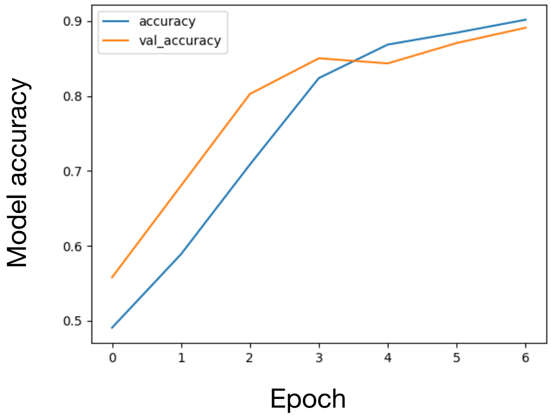
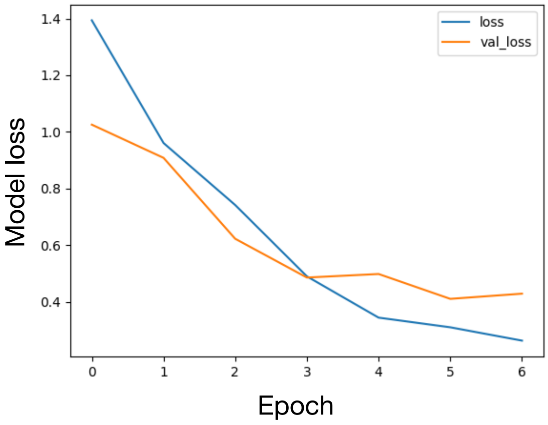


Fig. 3



We have evaluated the model using the unseen test set. The model has 80% accuracy with macro average precision and recall of 78% and 75%, respectively. The model has no trouble classifying normal and bacterial classes, however, its precision and recall of viral pneumonia class are significantly below average. It recalls only 56% of the viral class with 67% precision. This can be explained by the viral class having the least support. This conclusion is supported by the confusion matrix [Fig.4], which shows that most misclassifications happen (10) when viral (2) is classified as bacterial (1). Therefore, this can be improved by creating more images of the viral class using ImageDataGenerator or under-sampling the majority class and retraining the model.

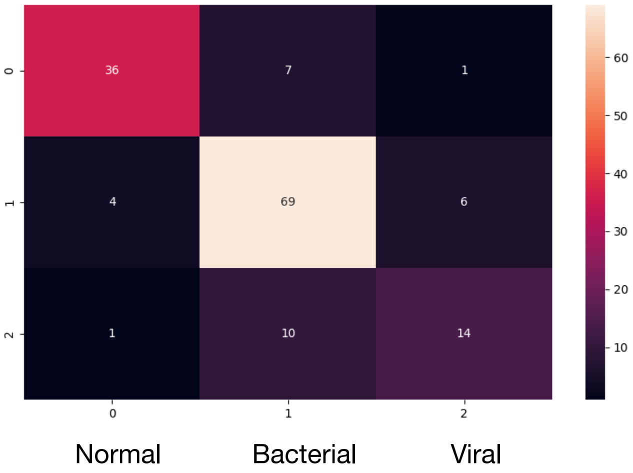
Tests

We have run the following preprocessing tests:

1. Distribution test — testing whether train, validation and test set split was completed correctly
2. Normalisation test — testing whether the data were normalised as expected
3. Dimension test — testing whether

Since some of the X-ray images were tilted and shifted from the centre, we have tested model resilience to noise and rotation. We have applied ImageDataGenerator to the test set to create a noisier set of images with up to 10 degrees of additional rotation, width and height shifts of 2.5% and 5% zoom.

Fig. 4 Test set confusion matrix



When tested on the noisy dataset, model accuracy dropped to 74%. We estimate that any further increase in noise will drop accuracy below 70%. In the future, we can increase model resilience by creating a noisy dataset and retraining the model on it.

Assumptions

Throughout the development of the model we have make the following assumptions:

- All the images in the provided dataset have been retrieved and resized correctly
- All input images are at least 256 by 256
- All images are grayscale
- All the images belong to one of the three classes — normal, bacterial, viral
- The bodies in the images are rotated no further than approximately 20% degrees
- No images were horizontally flipped
- Images on which the model will be used will have no significant zoom (above 10% with respect to the dataset)