BIRMINGHAM CITY
University

CMP6228 - Deep Learning Project

Project Proposal

Lewis Higgins - Student ID 22133848

CMP6228 - Deep Neural Networks

Module Coordinator: Khalid Ismail

Word count excluding figures, references and appendices: XXXX / 1500

Contents

Introduction

In this report, a novel solution is proposed to address a significant data science problem in the medical field, in the form of a deep neural network to accurately identify the presence of pneumonia from an image of a chest X-ray. To do so, this neural network will be trained on a publicly available dataset that has been previously seen across many publications, and advanced techniques for the model will be discussed.

This proposal will specifically cover the motivation behind this project before exploring related literature and previous works in great depth. To conclude, an optimal model will be proposed based on the knowledge extracted from these related works.

Motivation and objectives

1.1 Subject area

Pneumonia is a lower respiratory tract infection (LRTI) commonly caused by viruses or bacteria wherein the alveoli of the lungs become clogged with pus and fluid, which can be life-threatening in people of any age, but especially in children and the elderly (NHS, 2017). The World Health Organisation (WHO) state that pneumonia is the single largest infectious cause of death in children, killing 808,000 under the age of 5 in 2017 (WHO, 2025). Even if pneumonia is survived during the initial infection, Allinson et al. (2023) write that those who contract the condition as a child are 93% more likely to die from respiratory diseases later in life.

It is therefore imperative that recent technological advancements are leveraged for the quick diagnosis of the infection to allow swift treatment to avoid life-threatening consequences.

1.2 Dataset choice

The chosen dataset is sourced from the Mendeley data repository (Mendeley Data, 2025), uploaded and created by Kermany et al. (2018, p.1127) in their research of the applications of neural networks for medical diagnoses[1]. The dataset contains 5,856 images of chest X-ray scans of children taken from the University of California San Diego in America as well as the Guangzhou Women and Children's Medical Center in China, and is 1.18GB in size. There are only two classes of images: those with pneumonia and those without, as depicted by Figures 1.1 and 1.2.

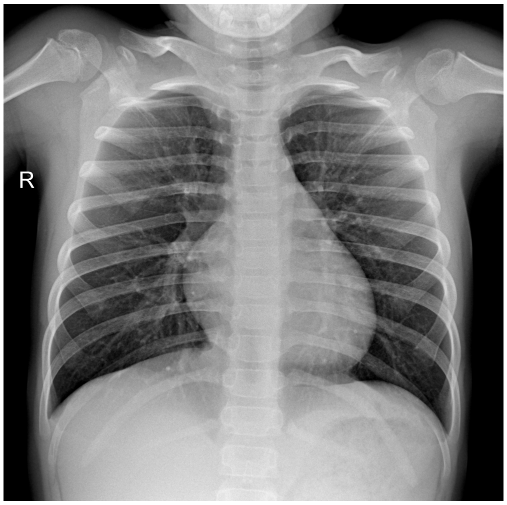

Figure 1.1: A sample image without pneumonia.

[1] Because their work was not only on pneumonia, the Mendeley ZIP file contains two separate datasets. This proposal is only for the "chest_xray" dataset.

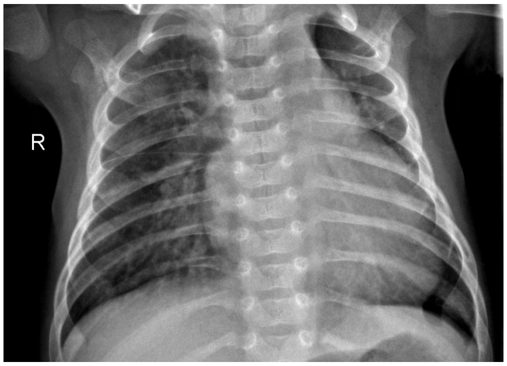

Figure 1.2: A sample image with pneumonia.

1.2.1 Additional observations

The dataset has already been pre-emptively split into training and testing sets, which saves some preprocessing. However, the data itself is imperfect and will require further preprocessing before being used to train a model. Table 1.1 denotes the potential issues with the dataset's initial state.

Issue	Explanation
Images are not all the same resolution.	The neural network's input layer will be fixed in size, meaning all input data must be the same size or the network will be unable to process it. This can be addressed using Keras to automatically resize all images to a given resolution.
Class imbalance	The training set contains 1,349 samples of patients without pneumonia, but 3,883 samples of patients with pneumonia. This will lead to the model favouring those with pneumonia rather than the underrepresented class of those without. This can be addressed using the techniques discussed in Appendix A.

Table 1.1: The issues with the dataset before any preprocessing.

1.3 Data science problem

This dataset poses a clear data science problem pertaining to the classification of these images which will be addressed through the development of a neural network image classification model leveraging supervised learning. The ground truth is already present within the dataset through its file structure, shown below:

```
.
‘-- chest_xray/
    |-- test/
    |   |-- NORMAL/
    |   |   |-- NORMAL-4512-0001.jpeg
    |   |   |-- NORMAL-11419-0001.jpeg
    |   |   ‘-- ...234 more images
    |   ‘-- PNEUMONIA/
    |       |-- BACTERIA-40699-0001.jpeg
    |       |-- BACTERIA-227418-0001.jpeg
    |       ‘-- ...388 more images
    ‘-- train/
        |-- NORMAL/
        |   |-- NORMAL-28501-0001.jpeg
        |   |-- NORMAL-32326-0001.jpeg
        |   ‘-- ...1347 more images
        ‘-- PNEUMONIA/
            |-- BACTERIA-7422-0001.jpeg
            |-- BACTERIA-30629-0001.jpeg
            ‘-- ...3881 more images
```

Files of the appropriate class are stored in the relevant subfolder. When this dataset is loaded, it will be possible to assign the relevant label to each image based on its subfolder of origin.

Related work

2.1 Introduction

"This section should demonstrate the main concepts of related techniques that have been previously used to solve the problem." What have other people done to solve it? How did they do it?

2.2 Lit Review Topic 1

2.3 Lit Review Topic 2

2.4 Lit Review Topic 3

Proposed model

"This section should demonstrate the suitability of the proposed solution in solving the data science problem"

Bibliography

Allinson, James Peter, Nishi Chaturvedi, Andrew Wong, Imran Shah, Gavin Christopher Donaldson, Jadwiga Anna Wedzicha and Rebecca Hardy (8th Apr. 2023). 'Early Childhood Lower Respiratory Tract Infection and Premature Adult Death from Respiratory Disease in Great Britain: A National Birth Cohort Study'. In: *The Lancet* 401 (10383), pp. 1183–1193. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(23)00131-9. pmid: 36898396.

Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia and Kang Zhang (22nd Feb. 2018). 'Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning'. In: *Cell* 172 (5), 1122–1131.e9. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.02.010. pmid: 29474911.

Mendeley Data (2025). *Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images - Mendeley Data*. URL: https://data.mendeley.com/datasets/rscbjbr9sj/3 (visited on 27/02/2025).

NHS (23rd Oct. 2017). *Pneumonia*. nhs.uk. URL: https://www.nhs.uk/conditions/pneumonia/ (visited on 27/02/2025).

WHO (2025). *Pneumonia*. URL: https://www.who.int/health-topics/pneumonia (visited on 27/02/2025).