

Analysing Pneumonia Disease Depending on X-Ray Images of Chest Using Deep Learning

تحليل مرض الالتهاب الرئوي اعتماداً على صور الأشعة السينية للصدر
باستخدام التعلم العميق

by

AHMED ADNAN KHAMEES

Dissertation submitted in partial fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

October 2023

Declaration

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the contents of this dissertation for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



Signature of the student

Copyright and Information to Users

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and as an open source and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

Abstract

Using Machine Learning (ML) in industry has vast applications, however using it in medical domain alerts a priority to help doctors determine unseen or hidden indicators of any probable illness or medical condition, which if not treated urgently may affect patient health. In this paper, the author aims to review and enhance Image recognition and classification using ML methodologies. The data input of X-ray images taken for medical proposes, used to gain better outcomes through advanced analysis of the training data, this includes specifying the average amount of data needed for training to make a good enough predictions using deep learning (DL) in order to save costs. In addition, exploring training data by applying data cleaning techniques to gain a well-balanced model for classification purposes. Author shown that setting 1600 x-ray images or more, as a training data input, tend to enforce a steady percentage of accuracy greater than 90%. Moreover, author described the results of using dirty (unclean) or unbalanced data to the ML model, which showed a clearly drop in precision, recall and F1 score percentages. Overall, our proposed experiments showed the importance of having a quality training data in achieving higher performance results.

Keywords: deep learning, convolutional neural network, image classification, x-ray, pneumonia

خلاصة

إن استخدام التعلم الآلي (ML) في الصناعة له تطبيقات واسعة، ولكن استخدامه في المجال الطبي ينبه الأطباء لوضع أولوية تساعد في تحديد المؤشرات غير المرئية أو المخفية لأي مرض أو حالة طبية محتملة، والتي إذا لم يتم علاجها بشكل عاجل قد تؤثر على صحة المريض. يهدف المؤلف في هذا البحث إلى مراجعة وتعزيز التعرف على الصور وتصنيفها باستخدام أساليب تعلم الآلة. يتم اختيار بيانات الصور الطبية الملتقطة عن طريق الأشعة السينية، والتي تستخدم بهدف الحصول على نتائج أفضل من خلال التحليل المتقدم لبيانات التدريب على تعلم الآلة، وهذا يشمل تحديد متوسط كمية البيانات اللازمة للتدريب لعمل تنبؤات جيدة بما فيه الكفاية باستخدام التعلم العميق (DL) في أجل توفير التكاليف. بالإضافة إلى ذلك، إن استكشاف بيانات تدريب تعلم الآلة من خلال تطبيق تقنيات تنظيف البيانات يعد أمراً مهماً للحصول على نموذج متوازن لأغراض التصنيف. أوضح المؤلف أن ضبط 1600 صورة أشعة سينية أو أكثر، كبيانات تدريب مدخلة، يميل إلى فرض نسبة ثابتة من الدقة تزيد عن 90%. بالإضافة إلى ذلك، أوصف المؤلف نتائج استخدام البيانات المتسخة (غير النظيفة) أو غير المتوازنة في نموذج تعلم الآلة، بأنها أظهرت انخفاضاً واضحاً في نسب الدقة والاستدعاء والنتيجة الكاملة لكفاءة ودقة النموذج. بشكل عام، أظهرت تجاربنا المقترحة أهمية وجود بيانات تدريبية عالية الجودة في تحقيق نتائج أداء أعلى.

Dedication

I dedicate this thesis to the following people who have helped me throughout my master journey:

To my dearest Mom, Wessam, who always remembered me in her prayers, you have taught me to be unique and determined. To my role model, my dad, Adnan, who build the man I am now and taught me to believe in myself. I quote from him, “the best investment in ourselves, is continuous learning”,

To my lovely wife, Isra’a, thank you for your many sacrifices. You have carried a double load, and you have done so without complaint. I will never forget your contributions; I could not have made it this far without your endless care, love, and support. To my little man, Wessam, who makes me happy for no reason and his innocence smile makes all the difficulties become much easier.

To my brothers and sister, Mohammed, Nawzat and Ala’a, whom their encouragement across many miles, and their longing to see me achieving my master degree, made the long days worthwhile,

Finally, I would like to dedicate this work to all of our families in Palestine, you are always in our heart, mind, and prayers,
I hope I have made you all proud.

Acknowledgements

First, and most importantly, praise and thanks be to Allah (God), the merciful, for giving me the inspiration, wisdom, strength, and endurance to complete this dissertation.

I would like to thank my supervisor, Professor Khaled Shaalan, for providing guidance and feedback throughout my dissertation. I appreciate his respect to his students, supporting and motivating them for more achievements.

Thank you The British University in Dubai for the happy memories, it has been such a nice journey throughout this master program.

Table of Contents

Chapter 1: Introduction.....	1
Chapter 2: Literature Review	5
2.1 Introduction	5
2.2 Review Related Work	6
2.2.1 Review Types of X-ray Images	6
2.2.2 Review Deep CNN Methods	8
2.2.3 Review Datasets Characteristics.....	17
2.2.4 Review Tools and Hardware Specifications	19
2.3 Review Conclusion	20
Chapter 3: Research Explanatory Methodology.....	21
3.1 Introduction	21
3.2 Scope of Research	21
3.3 Sample and Case Definition	22
3.4 Explanatory Research Hypothesis.....	24
3.5 Dataset Building.....	24
3.5.1 Data Collection.....	24
3.5.2 Data Analysis.....	25
3.6 Review Model and Research Questions.....	29
3.6.1 Model Training – First Research Question.....	29
3.6.2 Model Training – Second Research Question	30
3.6.3 Model Training - Risks and Proposed Solution.....	30
3.6.4 Model Training - Proposed Solution	31
3.7 Model Evaluation	31
3.8 Testing Hypothesis	32

3.9 Model Procedure and Tools	32
Chapter 4: Experimental Results.....	34
4.1 RQ1 Experimental Results	34
4.1.1 RQ1.1 Experimental Results	34
4.1.2 RQ1.2 Experimental Results	35
4.1.3 RQ1.3 Experimental Results	36
4.2 RQ2 Experimental Results	37
4.2.1 RQ2.1 Experimental Results	38
4.2.2 RQ2.2 Experimental Results	39
4.2.3 RQ2.3 Experimental Results	40
4.2.4 RQ2.4 Experimental Results	42
Chapter 5: Summary and Discussion.....	44
5.1 RQ1 Results Summary	44
5.2 RQ2 Results Summary	47
Chapter 6: Conclusion	50
Chapter 7: Future Work.....	51
References	52

List of Tables

Table 1: Reviewed Papers Data Collection..... 5

Table 2: Class Specific Data Distribution. 41

Table 3: RQ1 Model Efficiency..... 46

Table 4: Performance Summary..... 48

List of Figures

Figure 1: The Annual Size of the Global Data Sphere Projections from 2010 To 2025.	2
Figure 2: The Cost Using Technology Come Down Considerably Over Time.	3
Figure 3: Method for Analyzing Case Studies that Provide Explanatory Research Methodology.	22
Figure 4: Pneumonia Prediction Model.	23
Figure 5: Proposed Model Illustration.	23
Figure 6: Normal Chest X-Ray Image (No Symptoms).	25
Figure 7: Chest X-Ray Image with Pneumonia Symptoms.	26
Figure 8: X-Ray with Pneumonia and How it Looks when Processed into Numerical Representation.	27
Figure 9: Accuracy Curve.	35
Figure 10: Loss Curve.	36
Figure 11: Run Time Curve.	37
Figure 12: RQ2.1 Data Split.	39
Figure 13: Precision and Recall Results In Percentage.	39
Figure 14: RQ2.2 Data Split.	40
Figure 15: RQ2.2 Precision and Recall Results in Percentage.	40
Figure 16: RQ2.3 Data Split.	41
Figure 17: RQ2.3 Precision and Recall Results in Percentage.	42

Figure 18: RQ2.3 Data Split..... 43

Figure 19: RQ2.4 Precision and Recall Results in Percentage. 43

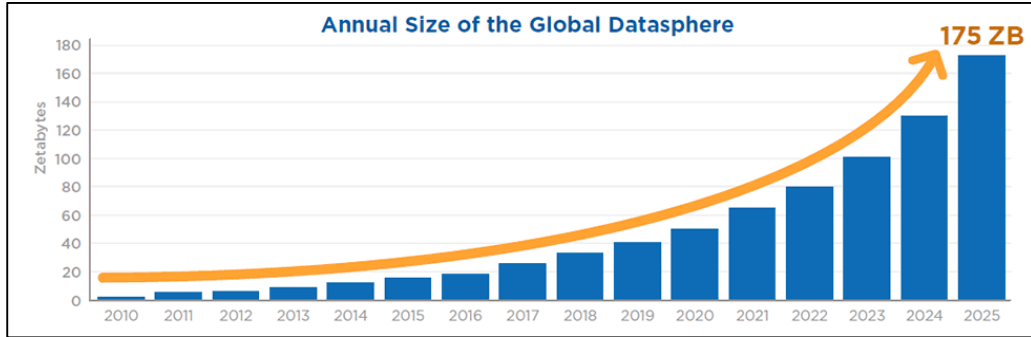
Chapter 1 : Introduction

As per [1] the World Health Organization, Pneumonia is defined as one of the acute respiratory type of infection that appears in lungs. When a healthy individual breathes, little sacs called alveoli in their lungs fill with air. Pneumonia causes the alveoli to fill with pus and fluid, which restricts oxygen intake and causes breathing difficulty. For children, pneumonia is considered as the major cause of mortality around the world. Pneumonia claimed the lives of 740180 children under the age of five in 2019, making up 14% of all pediatric deaths but 22% of all pediatric deaths between the ages of one and five. Pneumonia affects children and families worldwide, although southern Asia and Sub-Saharan Africa have the greatest mortality rates. Pneumonia can be avoided in children by using basic treatments, and it can be treated with low-cost, low-tech medication and care.

DL is used in many of state of the art business applications. It can process a large amount of data and learn to improve performance of a model against targeted results. Using Artificial Intelligence AI applications like Alexa, Siri, and facial recognition applications is becoming more relevant nowadays as an example of DL implementations, where it is applied to improve the performance using huge amount of data. Figure (1), a study by [2] shows the annual size of the global datasphere projections from 2010 to 2025, and the reading is around 175 ZB of information. This huge amount of data is hard for Humans to process. However using DL it is becoming much easier through time specially getting the benefit of advances in computer processing technology, that is wide available whether on cloud or premise.

Figure 1

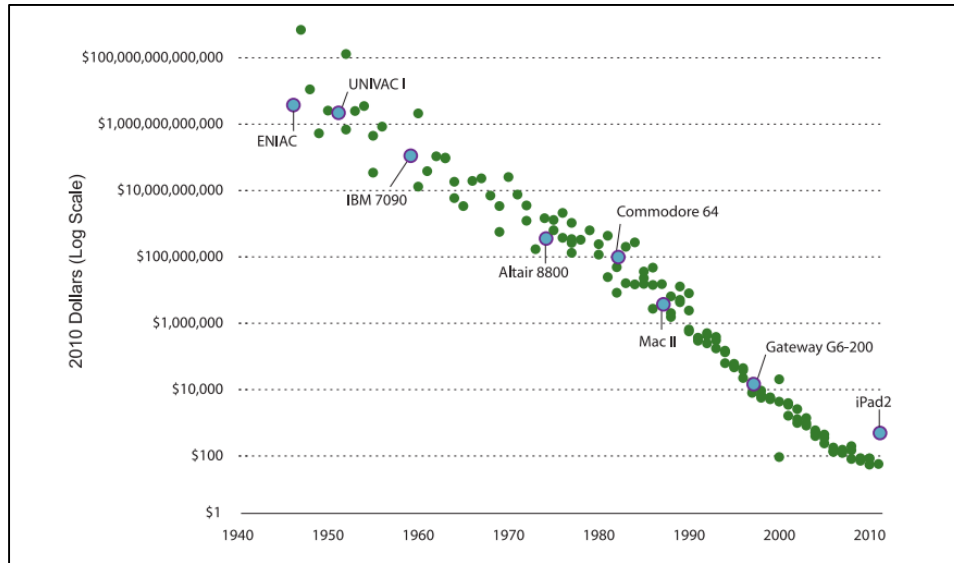
The Annual Size of the Global Data Sphere Projections from 2010 to 2025.



In addition, the cost of accessing to this computing power is dramatically decreased considerably which makes being able to deploy algorithms and run large scale of computations required for DL technique a lot more accessible to more people yet cheaper because of the high competency in the market. Figure (2) shows the cost using technology come down considerably over time that means running more DL learning computations using recent technology is easier nowadays [3].

Figure 2

The Cost Using Technology Come Down Considerably Over Time



Thus, using AI technologies especially DL, has reached to a point where it has a clear path to influence the current industries towards advanced transformation. Retail, marketing, sales and promotions, customer service management, supply chain management, and assessing risks, are examples of the industries that have great potential of innovation. Reducing processing cost of these tasks reaching out to targeted consumers, or reducing response time needed to handle some of customer requests, or avoiding any potential risks that may happen in future, all tend to support different industries for better decision making in advance.

The same, AI is implemented in medical field using DL powered solutions, to help specialists (i.e. doctors, nurses and caregivers... etc.) in their examinations to determine any possible risks of (e.g. tooth cavity, breast cancer, or Lung pneumonia... etc.), in a way that improves treatment procedure and patient health and experience. Using DL models over medical data is important to uncover unseen illness and help specialists to get additional useful insights about the case under investigation. [4]

In this paper, author proposed a solution using machine-learning techniques that depends on supervised learning classification.

In the following sections, author made literature review on the most recent image classification studies using deep learning algorithms, specifically those undertaken to analyse x-ray images. After that, author presented Explanatory Research (ER) been adopted, and how ER been used to evaluate the proposed DL model showing the experimental results, discussing them, and finally make the conclusion.

Chapter 2 : Literature Review

2.1 Introduction

Many researches proposed the using images for classification using DL. In this section, author investigates other researchers work that is related to our topic of interest in this paper (i.e. Medical images classification). The literature review demonstrates different medical field use cases that investigates using ML models for classifying different types of x-ray images. Author found many papers related to this topic as shown in table (1).

Table 1

Reviewed Papers Data Collection

Serial	Literature Review papers	Collected Data
1.	Search Topic	Medical images classification Recent papers: since 2018.
2.	Search Boundaries	Papers with applied classification models and results.
3.	Needed (included)	24 papers
4.	Relevant: Not Needed (excluded)	20 (e.g. reviews)
5.	Not Relevant Papers	53 Papers
6.	Total Number of investigated papers	97 papers

While there is a broad range of theories covered in the literature, this analysis will concentrate on four key elements that recur often in the review. Firstly, author goes through the specifications of different X-ray images in the medical field. Then, Author reviewed the

various techniques of DL models used with different types of Convolutional neural networks (CNNs). After that, author looks at the dataset characteristics used in DL analysis (i.e. its origin, type and size). Finally, author shows what computer specifications was preferred to use in for processing CNNs and conducting experiments.

2.2 Review Related Work

2.2.1 Review Types of X-ray Images

Deep learning applied in wide range of clinical departments, [5] applied it into two applications, first one is using 2D segmentation in annotating training set of two types of organs using fetal magnetic resonance slices, and second one is using 3D segmentation in annotating tumor core from different brain tumor MR sequences. [6] proposed a method that classifies patient's medical images (lung, brain, breast cancer) into either benign (negative cancer) or malignant (positive cancer). [7] authors used ImageCLEF 2015 and 2016 datasets. 2015 version consists of different categories of medical cases (e.g. CT and MRI) with 6776 images (i.e. 4532 for training and 2244 for testing). Whereas 2016 version consists of total 10942 images (i.e. 6776 for training and 4166 for testing) images. On the other hand, ISIC 2016 and 2017 datasets represents medical images to classify melanoma diagnostics as either benign or malignant with total 1279 images (i.e. 900 for training and 379 for testing). Whereas 2017 version consists of 2750 images (i.e. 2000 for training, 150 for validation, and 600 for testing) to classify nevus, melanoma and seborrheic keratosis vs others. [8] This model is based on training 4298 3D images of five anatomical categories. Images were mined from a set of 1675 patients from a hospital archive. Classified images original resolution is 512x512 and rescaled to the half (i.e. 256x256).

[9] adopted 37698 images for evaluating the models, 37198 of them used for training the CNN architectures and 500 images for testing. The experiment considered five anatomies (i.e. heart, liver, kidney, lung and lumbar spine. [14] authors obtained image dataset of biopsies patched into total 91899 images that contains the three classes under study (i.e. CD, EE, and normal ones) distributed into 30198, 29308, and 32393 patches respectively. Images gathered from 150 children with median 37.5 age. The size of image patches in the input layer comes with 1000x1000 pixels. [15] authors suggested methodology used a brain MRI images brought from a local source. Images are rescaled to size of 227x227 pixels. [16] used labelled optical CT images and X-ray for chest dataset for conducting experiments. The Chest dataset consists of 5856 images with 8-bit grayscale. The size of this dataset are multi-sizes ranging between 384×127 pixels and 2916×2583 pixels. [18] used two datasets that represents four different tissue types. Images size is 720×480 pixels. ISIC2017 dataset represents skin lesions acquired from the International Skin Imaging Collaboration (ISIC) and consists of 2000 medical images of skin tumors (Melanoma 374 image, and Nevus of Seborrheic Keratosis 1626 image). The other dataset HIS2828, consists of 2828 images of several tissues (nervous 1026, connective 484, epithelial 804, muscular 514). [20] authors used images of CT and X-Ray images that has 6087 total images of normal (1583) and pneumonia (4504). All dataset images size was rescaled to 224x224 or to 299x299 pixels. [22] authors used x-ray dataset of COVID-19 infected patients and normal images from a public dataset found on kaggle.com. Images was reset to a standard size and normalized to 224x224 pixels. These x-ray images also been augmented before inserting them for training by scaling, translation, and rotating. [23] authors used utilized dataset for training the proposed model and it consists of total 3905 x-ray images that corresponds to six types of pulmonary infections besides COVID-19 disease. Images were rescaled to avoid distortion and different pixels ratios to 200×200 pixels

size. Moreover, in order to make a complete transformation authors added a 200 x 200 pixels black background and excluded thoracic x-ray images that has low contrast. Augmentations were slightly applied during training phase which is mandatory to crucial for aiding CNNs capabilities. [24] authors applied their proposed method of classifying brain tumor on two dimensional (2D) image CE MRI dataset that includes three categories of brain tumors. Each image with size 512 x 512 pixels. [25] Authors used images of Digital Mammogram dataset to help in breast cancer diagnosis. Images has 512 x 512 pixel resolution, and $32 \times 32 \times 1024$ feature map.

2.2.2 Review Deep CNN Methods

Convolutional neural networks (CNNs) attained promising outcomes in terms of performance and accuracy. However, each CNNs models implemented deep learning in different approaches and steps. In medical sector, (CNNs) attained good performance in segmenting images automatically, but still not enough accurate for the use in a wide range in clinical field. Moreover, using interactive segmentation over CNNs faced other challenging matters, where recent CNNs cannot do enough generalization to objects that are previously unseen and not shown in the training dataset. Thus, [5] authors designed a flexible CNN framework with 2D and 3D interactive segmentation for medical images used as an input for this deep learning based method. However, to avoid overfitting authors of this framework proposed special image fine-tuning for that makes the CNN flexible to either supervised or non-supervised test images, yet leads to more accurate outcomes in faster manner than it is in state of the mind similar methods. [6] Authors proposed Optimal Deep learning (ODL) a new enhanced classifier for lung, brain, breast cancer, and Alzheimer illness images. Some of these disease images are small in size that makes it hard for the classifier to detect it accurately. To

overcome this issue, the introduced method used Histogram Equalization to select optimal features (e.g. grey and multi-texture features) using a new algorithm based on Opposition Crow Search (OCS) in the pre-processing phase, this will make images better in quality as an input for classifier. As a result, ODL classifier increased results diagnosing any illness in medical images with 92.23% for accuracy, 86.45% for sensitivity and 100% for specificity. [7]

Authors presented synergic deep learning (SDL) model that uses multiple (n) CNNs simultaneously, this will enable them to learn mutually from one another. Where each pair predicts if the images are from the same type of class. Moreover, the model learns unique features from pair of images in the same time simultaneously. The introduced model evaluated on four datasets of medical images. All this helped SDL model gain better results and made extra improvements in classifying medical images achieving state of the art performance. [8]

authors presented a supervised learning method to classify specific anatomical body parts using images captured by computed tomography (CT). Data augmentation been applied in order to improve the performance of the classifier to capture the essence of dataset images in better manner than it is if the training be applied using hand crafted features. This model achieved specific error rate of 5.9% and average testing values of the area under the curve equal to 99.8%. This is crucial for training images that may help in medical diagnosis. [9]

authors proposed a new modified and robust CNN architecture that classify medical images automatically to distinguish anatomical features by doing learning at multi-levels of abstractions of dataset. This model is compared to other three key architectures (i.e. GoogleNet, AlexNet, and LeNet). These three architectures overfit because the hyper parameters and layers number is used over a large dataset of images. Based on comparative evaluation, experiments shown that the proposed CNN outperforms the other three architectures in anatomical objects classification of medical images dataset, by achieving 81%

in accuracy, next AlexNet have had only 74%, then LeNet got 59% and finally GoogleLeNet achieved 45%. [10] proposed a deep CNN model that classifies breast cancer histopathological images as benign or malignant classes. The used model is from Google (i.e. V3 inception model) that is trained to classify non-medical images. The model is based on transfer learning and data augmentation that employs images over a pre-trained model. The goal is to investigate feasibility of using the proposed model in classifying medical images related to detection of breast cancer. The model achieved 83% accuracy in classifying benign and 89% for classifying malignant. [11] Authors introduced a synergic deep learning (SDL) model based on a dual deep CNN. This model works as mutually designed signal system used for input image verification whether the input pair of images are belonging for the same class, then make a correction if an error shown. This model especially designed for confusing input samples to solve classification problems of medical images. Experiments of the proposed SDL over ImageCLEF2016 dataset shown better results in terms of performance and accuracy than it is achieved by the winner solution of Subfigure Classification Challenge. [12] authors examined alternative strategy called Enhance-Net to increase the performance of deep learning (DL). The introduced model consists of four main stages where data is prepressed then trained and validated using a benchmark DL model called Champion-Net. The main characteristic of the proposed model is how image been enhanced to improve the outcomes and performance. Enhance-Net validation is used to measure the performance against lightness order error (LoE). Radiologists verified the model using real time training and testing techniques and achieved 97.74% in accuracy. Thus using the proposed model Enhance-Net improved DLMs over musculoskeletal radiograph-bone (MURA-BC) dataset significantly for clinical use of x-ray images. [13] authors presented in their study GoogLeNet CNN classification model for categorising axial CT images into different classes of anatomy based on their dataset training

size (i.e. 200, 100, 50, 20, 10 and 5) then images tested the model with 6000 CT images. The goal is achieving better accuracy results and lower variance when using CNN classification model with optimum size of medical images training data and lower standard deviation.

Experiments been conducted over dataset brought from Picture Archiving and Communication System (PACS) in Massachusetts General Hospital (MGH). Results shown that the accuracy of the classification model increased expeditiously from training images with size 5 until 50. However, the accuracy did not increased typically with training images of size 100 until 200.

[14] Authors introduced a new hierarchical approach for classifying medical images that is based on using multiple deep CNNs. For hierarchically organized medical images, it is hard to apply multiclass approach using classical supervised methods of deep learning. Thus, the new approach introduce a new way to deal with medical images hierarchal levels as stacks for better comprehension. Here, Authors used CNNs hierarchical model that get results based on two levels for experimental purposes. The experiment applied on a biopsy image from the small bowel to distinguish different kind of disease from parents and child levels. In the pre-processing phase, images are went through patching, after that clustering images to remove unnecessary information, then colour balancing is done in order to fix any staining issues. Results of testing on this dataset of medical images shown that this process yields to robust outcomes at both lower and higher levels. Moreover, accuracy was continually better than other obtainable CNNs, DCNN and multi-layer techniques. Thus, using hierarchy in deep learning shows that results could have classification improvements and flexibility when data is classified based on hierarchy over other traditional techniques that consider only multiclass problems. [15] authors introduced a solution that helps in detecting brain tumor based on watershed algorithm then using CNNs. Brain images taken using Magnetic resonance imaging (MRI) technique are uploaded into a system for tumor detection. Firstly, images are

pre-processed by sharpening them into grey scale level. After that, thresholding is applied for segmentation, and then removing noise from the original image. Next, the MRI image of the brain is classified whether it is tumor or not using CNN algorithm based on the size and location of the determined tumor. Identifying the damaged of the brain is done using a marker, and tumor shape is detected as well from the input MRI scan image. Results shown high outcomes with 98% accuracy. [16] authors introduced an adaptive DL model to solve the challenging issue of classifying different sizes of medical images. Solution been applied to classify pneumonia CT dataset of medical images into two binary classification types (i.e. pneumonia vs. normal). Comparing to traditional models (i.e. VGGNet19, VGGNet16, AlexNet and ResNet-50) the proposed work improved the classification by around 4% in terms of effectiveness and overall performance. Moreover, it is more practical for classification of medical images with multi-size. [17] authors built a DL algorithm for detecting oral cancer in patients by analysing their hyperspectral images based on bagging and boosting methods. The structure is based on deep CNN partitioned into two layers for setting labels and classification through highlighting the area of interest of the hyperspectral image. Validating medical images classification performance is done in terms of specificity, sensitivity, and accuracy by comparing the proposed method with other partitioned DL methods (i.e. DBM and SVM). Obtained results for conducting the proposed algorithm is 0.91 for specificity, 0.94 for sensitivity, and 91.4% for accuracy. Thus, results showed increasing in overall outcome for diagnosis of oral cancer by classifying medical images using partitioned CNN model. [18] authors proposed a new DL model for classifying medical images, the classification algorithm is based on Combining Network together with Multilayer Perceptron (CNMP), where some of traditional features are extracted along with high level features from deep CNN. Results shown that the proposed classifier achieved better results in terms of

accuracy, here where two datasets of medical image are used (ISIC2017 and HIS2828).

ISIC2017 obtained 90.2% accuracy, and HIS2828 obtained 90.1% accuracy. Thus, outcomes of CNMP considerably outperforms other models (i.e. coding network, traditional SVM, and R feature). [19] Authors proposed a study to evaluate the performance of several deep learning models based on CNNs for detecting and classifying objects by running exhaustive experiments on the newly gathered medical images of breast breast lesions. The experiment is applied in two steps. Firstly, is to locate the areas of interest where the lesion may occur. Secondly, to classify these highlighted areas if a benign occurred or not. For evaluation process the following architectures been explored (VGG, ZFNet, AlexNet, GoogLeNet , DenseNet, and ResNet). Results shown that CNN with deeper number of network and less number of parameters achieves better outcomes using the collected dataset. Moreover, using DenseNet is further applicable and suitable in comparison with other evaluated architectures. [20] authors conducted a an experiment to make comparison between different CNN architectures to classify binary images that detects pneumonia based on upgraded form of CNNs (VGG19, VGG16, Inception_V3, Inception_ResNet_V2, DenseNet201, Xception, Resnet50, and MobileNet_V2). L2-regularizers and weight decay were used to avoid any possible overfitting issues. Results shown that upgraded form of Resnet50 achieved very well outcomes in terms of performance with increasing rate of accuracy for training and testing (higher than 96%). [21] authors established a new tool for diagnosing diseases on patients with blinding retinal issues where vision may be lost because of that disease due to either a seen choroidal neovascularization neovascular AMD (Age-related macular degeneration) or because of macular edema for patients with diabetic. The tool is based on transfer learning neural network where training on data is done using conventional techniques. Results of demonstrated model shown competitive performance in analysing Optical CT images where

neither a special DL machine is not required nor a huge dataset number of images. Moreover, the classifying performance shown close outcomes in comparison with expert human with related experience in diagnosing retinal diseases. In addition, training models with less images i.e. around 1000) of each category, shown high steady performance in terms of specificity, sensitivity, and accuracy. Thus, applying transfer learning with less number of training data achieved classifications with higher effectiveness. [22] authors proposed deep CNN model to find COVID-19 in x-rays images of chest by distinguishing the normal healthy lungs against those has pneumonia. CNN models are pre-trained (InceptionV3, EfficientNetB0, and VGG16) using transfer-learning technique for model evaluation. These models proven efficient implementation and their balanced outcomes in terms of accuracy with less parameters the best work with applications that helps in COVID-19 detecting. Experimental results shown high efficient and quality model reported from the EfficientNetB0 with 94.79% accuracy in classifying and detecting COVID-19 infection from the provided chest x-rays of all other possible abnormalities, and shown 92.93% overall accuracy. Thus, it was proven that transfer learning is valuable in enhancing models' learning capabilities. Moreover, the high accuracy outcomes shown an important evidence that technology in computer vision applications can gain better accuracy results in detecting COVID-19 infection with fewer interventions from radiologists. [23] authors presented Mobile Net, a state-of-the-art CNN which is employed to investigate the important of feature extraction for medical images classification. The purpose of this model is to detect recently emerged pulmonary infections specifically COVID-19 by examining x-ray images. The employed model is performed to implement experiments in three steps for feature extraction inspection, by firstly training the model from scratch, then extracting features through transfer learning, then by fine-tuning. Results shown that implementing CNNs with training from scratch achieved higher

performance in comparison with other techniques of transfer learning. Where accuracy of classifying all seven x-ray pulmonary infections achieved 87.66%, whereas distinguishing if COVID-19 infection occurs or not achieved 99.18%, 97.36%, and 99.42% in terms of accuracy, sensitivity, and specificity respectively. [24] authors proposed a new method to classify images of brain tumor using fine-tuning approach called block-wise that is based on transfer learning technique. This approach suggests different path in comparison to implementing pre-trained CNN model, that train classification method separately such as (Support Vector Machines SVM, Decision Trees DT, Boosted Trees, and K nearest neighbours). Benchmark dataset was used for evaluating the proposed method T1-weighted contrast-enhanced CE magnetic resonance images MRIs. Results compared with other three DL CNNs methods and pre-trained using VGG16 and AlexNet achieved 94.65% and 89.95% respectively. Thus, VGG16 classification model outperformed other pre-trained models using CE MRI dataset. [25] Authors proposed a new model based enabling feature transfer with multitask deep learning (FT-MTL). MTL models can implement three tasks (detecting objects, segmenting, and classifying) jointly in which it saves cost of computing yet improves robustness to avoid overfitting issues but on the other hand, it fails to benefit of combining each feature in each task any earlier in the training stage. Thus, FT-MTL network utilizes these three tasks of same data sources known as domains by which features exploited from multiple domain can be beneficial in enhancing generalization of implementing the network more broadly. In FT-MTL network validity evaluation, authors compared it with literature that includes eight models of classification, four of detection, and three of segmentation using dataset of mammogram images of breast cancer where five-fold cross validation is applied for measuring the performance. Experiments shown that the proposed model outperforms other eight models in terms of classification achieving area under curve AUC performance metric

with $92\% \pm 0.02$, and outperforms detection task of other four methods with $91\% \pm 0.05$ true positive rate. However, in segmentation the model achieved comparable results of $76\% \pm 0.03$ as average dice index metric. [26] Authors applied CNNs model for detecting coronavirus disease by implementing it over chest CT images that may indicate the presence of this infectious disease. The model can learn features from the dataset of medical images and extract them automatically. COVID-CT dataset of 425 coronavirus CT images is used to train GoogleNet architecture with inspectionV1 model. The experiments shown promising results for training GoogleNet with 82.14% validation accuracy. Moreover, testing the model achieved 84.01% precision, 82.11% specificity, and 83.12% sensitivity. [27] Authors presented a CNN model used for detecting and classifying pneumonia infection in x-ray images of samples taken from patient's chest. The model is constructed from scratch to challenge other state of the art and traditional CNNs methods. The purpose is to do feature extraction from the given dataset and deploying algorithms the do data augmentation in order to validate and improve the accuracy of classification of the proposed model. Experiments achieved remarkable results in validation performance (0.1288 training loss, 0.9531 training accuracy, 0.1835 validation loss, and 0.9373 validation accuracy). [28] authors described building a 3D fully convolutional network (FCN) and how can 3D images be processed to automatically produce semantic segmentations. FCN was introduced to overcome the CNNs limitation of losing images spatial information in the final conv-net connected layer where features are input to it. In FCN, transposed conv-layers replaced the final conv-net of traditional CNN in order to apply samples learned up into the features with low resolution within targeted network. As a result, spatial dimensions are recovered into its original as input image while segmentation is performed semantically at the same time. The proposed model achieved promising performance results with 90% testing score. Moreover, authors shown that

using 3D deep FCNs over modern GPUs can be efficiently implemented for training the model.

2.2.3 Review Datasets Characteristics

Applying deep learning in medical sector is crucial for medical stuff to help them detecting and classifying defected areas that appears important and sensitive main parts of patients body (i.e. brain, chest, skin, liver, lungs, breast, oral, gastric and Celiac Disease) using different types of x-ray images . [5] Used training set from the 2015 Brain Tumor Segmentation Challenge (BRATS) that has 274 scan images taken from 198 patients. [6] Used four datasets in their classification method in order to get better accuracy in predicting the four type of diseases under investigation (i.e. Alzheimer, lung, brain, breast cancer). [7] evaluated their proposed SDL model over four image datasets, two of them were based on pathology, and been presented in the Cross Language Evaluation Forum 2015 and 2016. The other two image datasets were based in modality, and been presented in the International Skin Imaging Collaboration (ISIC) challenge for classifying skin lesion. [8] used medical images retrieved from archiving system in the national institute of health many of them related to the Liver disease. Then images labelled manually based on tags of examined body parts. [9] used a dataset brought from Department of Health and Human Service in the National Institutes of Health of the the U.S. National Library of Medicine. This dataset used to help computer aided systems in diagnosing different anatomical objects between patients (i.e. normal lung compared to damaged due to illness, also small objects of spine bone is different between individuals due to medical condition that may getting worse). [10] used a database called BreakHis that consists of 7909 images of both benign and malignant classes that detects breast cancer. This dataset of microscopic biopsy images was taken from 82 patients. [14] authors used medical images collected from patient's biopsy used for performance testing and

categorised into three classes for parent's level (i.e. Celiac Disease CD, Environmental Enteropathy EE, and histologically normal patches) on the other hand, on the child level, severity of CD is classified into four classes (IIIC, IIb, IIIa and I). Dataset were gathered from three universities (i.e. the University of Zambia, The University of Virginia, and the Aga Khan University Hospital in Pakistan).

[17] used partitioned CNN classifier with medical images to help in detecting oral cancer, hyperspectral images where from the following datasets: The Cancer Imaging Archive (TCIA), Genomic Data commons (GDC), and repository of BioGPS UCI for normal data extraction. [19] Authors gathered new dataset for detecting breast lesions, 1043 cases in total (i.e. 464 for malignant and 579 for benign) ultrasound images annotated manually by expert medicals from Sichuan Provincial People's Hospital.

[25] Used dataset published online publically in INbreast database that has 410 images belongs to 115 patients with several lesions (asymmetries, masses, calcifications, and distortions). INbreast is a full-field digital mammographic database is created by the Breast Center in CHJKS, Porto, where they took permission from both the National Committee and Hospital's Ethics Committee of Data Protection.

[27] Used a dataset that consists of three main folders for training, testing, and third one for validation. In addition, the dataset has two sub folders that contains 5856 total chest x-ray images of two types, one for normal, and the other one for pneumonia. Chest images belongs to paediatric patients whose age ranges between one to five years old. [28] authors used in their proposed FCN model a dataset brought from a research study done with patients who suffers from gastric cancer. The dataset includes scans of 377 abdominal CT images that were enhanced in terms of contrast. Every CT volume included of 460 to 1177 slices with size of 512 x 512 pixels.

2.2.4 Review Tools and Hardware Specifications

In order to process CNNs to achieve best performance outcomes in terms of accuracy and cost of implementation among others, authors used computer devices equipped with different specifications of both hardware and software. [9] authors conducted their experiments using a device integrated with NVIDIA GeForce GTX 980M. [10] used TensorFlow application to utilize learning transfer Google's concept of V3 inception model using python. [12] used a virtual environment to conduct the experimental research. The operating system is Ubuntu what has six (Intel Xeon 2.2 GHz) CPUs and 12 GB RAM were allocated. Simulation used Python 3.0 for processing. [13] Authors used NVIDIA DL GPU Training System (DIGITS) DevBox, four TITAN GPUs, memory bandwidth (336.5 GB/s) and 12 GB (memory/board) to train each dataset with GoogLeNet model. [16] authors conducted their experiments using Tensorflow application over Ubuntu System (16.04.5 LTS version). The hardware is equipped with NVIDIA GeForce (GTX 1080) version and AMD processor (i.e. Ryzen 5 1600) that has six-cores. [17] Authors implemented their experiment based on hardware with one TB hard disk, NVIDIA GeForce GPU, i7 Intel processor, and 64 GB RAM. [18] Implemented CNMP algorithm in a matlab toolbox called MatConvnet with hardware specification experiments on a computer with 32G memory, core i5 CPU with 3.2 GHz, and Graphics processing unit of GTX1060. [20] authors conducted their experiments using a hardware with that has 8GB RAM and Intel CPU Core i7-7700 3.6 GHz and running on Windows 10 64 bit. Tensorflow tools used for implantation of deep learning and NVIDIA (Tesla P40) GPU with (24 GB) RAM. [26] Authors used a computer that is equipped with Core i7 CPU, GeForce GTX 1070 GPU, and 16 GB RAM. In addition, the model been executed with windows 10 home edition and MATLAB 2018 version. [27] Authors implemented the proposed CNN model using a standard computer that has CUDA Toolkit 9.0

installed and equipped with 12 GB Nvidia GeForce GPU card accelerated with cuDNN v7.0 library.

2.3 Review Conclusion

From the previous review, it can be found that the general desire is to improve the performance results when using different DL methods. However, the central issue which researchers focuses is specifically when implementing DL over medical image datasets that occurs in different human body parts (e.g. lung, chest, brain, breast, skin, heart, liver, kidney, and lumbar spine). These reviews show how DL is important in detecting some critical diseases (e.g. brain cancer, melanoma, and pulmonary infections such as COVID-19), most importantly Pneumonia that is related to the topic of this research. Where most of the literature argues about using different versions of CNNs models and in some cases some upgraded ones (i.e. ODL, SDL, GoogleNet, AlexNet, and LeNet, VGG, ZFNet, DenseNet, and ResNet). Here comes the importance of this review, where author promoted the importance of maintaining quality training data, rather than investigating these different CNN models, for getting results that are more realistic, stable, and solid.

Chapter 3 : Research Explanatory Methodology

3.1 Introduction

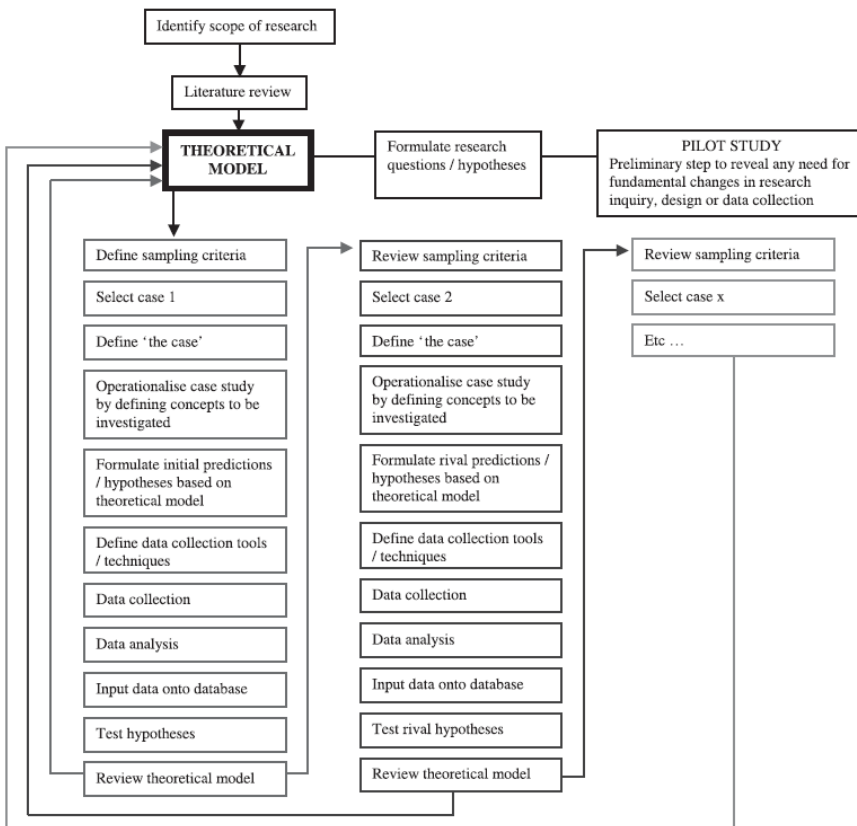
In general, author is interested in Explanatory Research (ER) to help in find explanation and answers of (how) questions and in understanding the problem phenomenon and its causes. In addition, ER is considered as one of the effective tools for making future predictions based on current observations. Thus, author considered ER being a great theoretical model to build the subject of this research.

3.2 Scope of Research

In this section, more specifically, author adopted [29] proposed explanatory methodology for clinical research analysis to answer the general research problem of this paper (i.e. Applying ML model to study how does the detection class might be impacted by using different quality versions of categorical training data?). Here where [XX] explanatory model has added on over traditional quantitative analysis, since the proposed explanatory model provides flexibility in evaluating complex multifactorial clinical practices as shown in Figure (1).

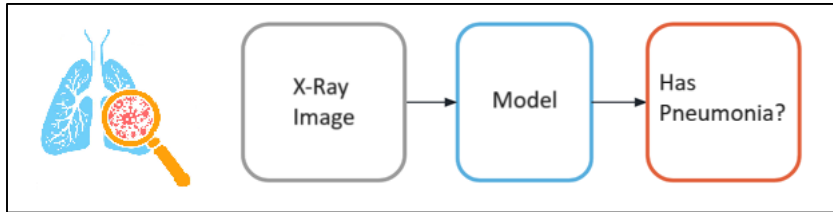
Figure 3

Method for Analyzing Case Studies that Provide Explanatory Research Methodology

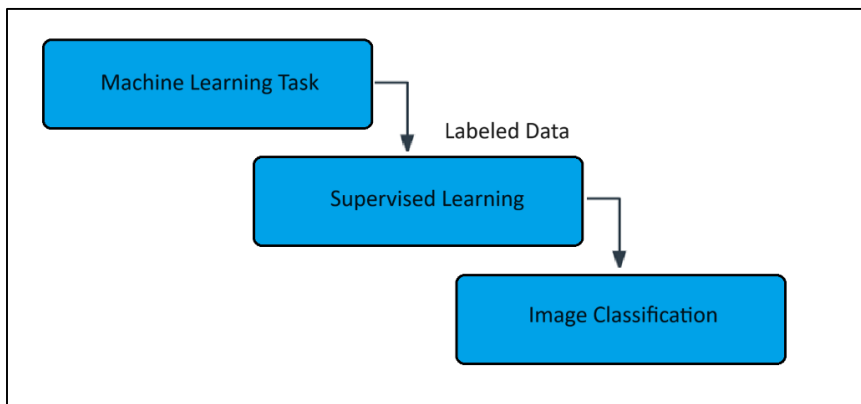


3.3 Sample and Case Definition

In this research, and as shown in figure (4), author aims to predict if x-ray Chest images belongs to one of the two x-ray samples (classes) Pneumonia and Normal.

Figure 4*Pneumonia Prediction Model*

Thus, one of the great ways to answer the research question of this case problem is machine learning. Author proposed a model provided with labeled data.

Figure 5*Proposed Model Illustration*

In this research, author proposed a solution for x-ray images classification task based on machine learning that is performed by supervised learning, as shown in figure (5). That means we selected a range of x-ray images that were labeled previously by medical experts, and refers to the categorical sets under investigation. Categorical labels have two discrete possible values, in our case either Pneumonia or not Pneumonia (Normal), where Pneumonia is divided into two subcategories Pneumonia caused by Bacteria, Pneumonia caused by Viral (e.g. viruses). In the first case study of this paper, Author used Pneumonia x-ray images caused by Bacteria and Viral for comparison against Normal. After that, Author used

Pneumonia caused by Bacteria and Viral separately to do held the second case study for advanced experimental analysis.

3.4 Explanatory Research Hypothesis

For the first case study of this paper, Author investigates whether if, keep adding more training data into the ML model, will increase the accuracy steadily or not. Therefore, the first hypothesis is:

H₁: Adding more training data to the ML model will increase Accuracy by 4%.

For the second case study of this paper, Author investigates whether if, using ML model with dirty (unclean) or unbalanced data or both will have an impact on results in terms of Precision and Recall. Therefore, the second hypothesis is:

H₂: Adding dirty (unclean) or unbalanced data to the ML model will clearly drop precision and recall results.

3.5 Dataset Building

3.5.1 Data Collection

Author explored different datasets in order to solve research task using ML. Understanding dataset composition is needed to help in building better ML models and more efficient solution to our task. Thus, choosing a high quality images is essential to any ML task especially in the medical field. In this research, the author used a dataset found on open source portal (kaggle)¹ contains several x-ray images that belongs to patients suffers from different chest diseases as follows: Pneumonia caused by bacteria, Pneumonia caused by virus, in addition to Normal cases.

3.5.2 Data Analysis

The dataset contains around 2GB of x-ray images (i.e. 5863 images of two categories Pneumonia and Normal). In our research, we used a subset of this dataset as follows (100 Pneumonia-bacteria, 100 Pneumonia-Virus, and 100 Normal).

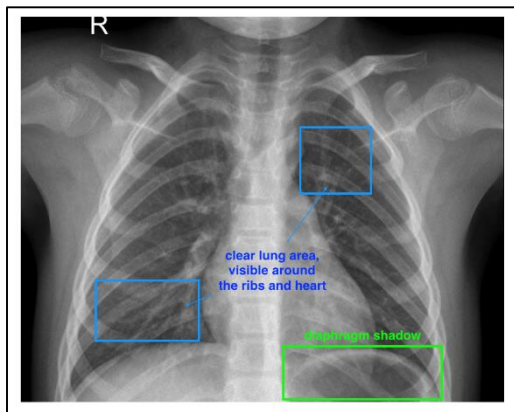
(1) <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>

3.5.2.1 Exploring Data

We can explore few symptoms that help in indicating pneumonia visually as shown in the following pictures.

Figure 6

Normal Chest X-ray Image (no symptoms)



Paying attention to lungs and diaphragm areas are most important to help radiologists in indicating any possible occurrences of pneumonia symptoms. Figure (6) shows a normal case of chest x-ray that depicts a clear photage of lungs and diaphragm without any abnormal opacity or cloudiness.

Figure 7

Chest X-ray Image with Pneumonia Symptoms

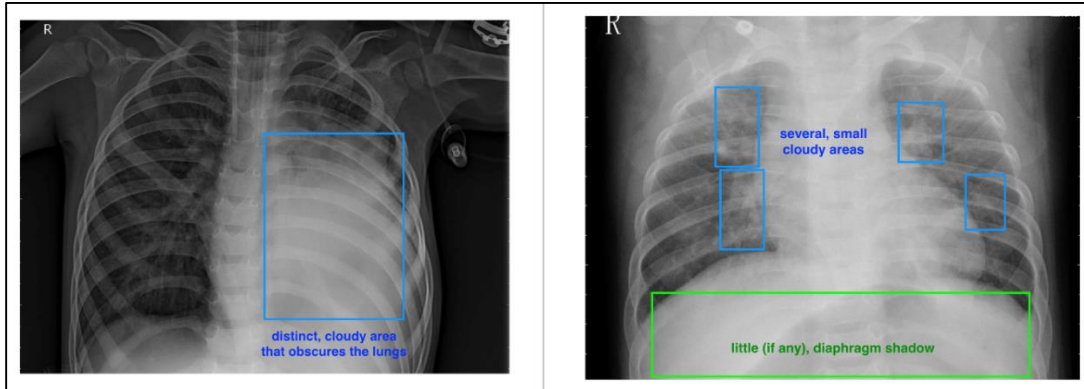


Figure (7) shows Chest x-ray image with Pneumonia symptoms; the left chest image shows opaque area concentrated over the right lung. The right chest image shows different areas of opacity and cloudiness in several concentrated parts of lungs, diaphragm, and heart.

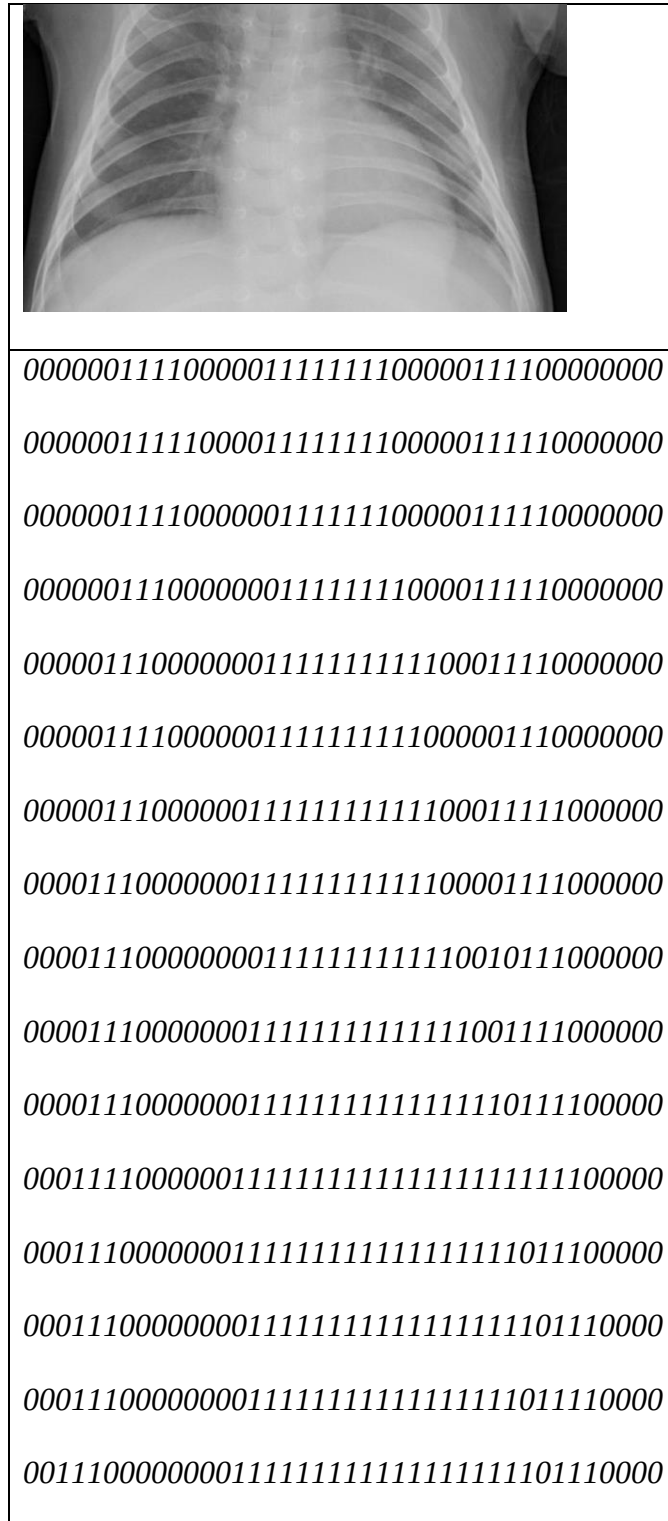
3.5.2.2 Data Vectorization

Our ML model requires the data to be numerical, so we transformed all of the chosen subset of data into numerical format. This is done automatically using python tools.

Presenting an image with 0 and 1 black and white. For example, figure (8) shows an x-ray image diagnosed to have Pneumonia in the white area. Below it, we can notice the processed image into numerical representation.

Figure 8

X-Ray with Pneumonia and How it Looks When Processed into Numerical Representation



3.6 Review Model and Research Questions

In this paper, author goal is to help radiologists in flagging serious cases of Pneumonia in children, and identify healthy normal cases more accurately, and, generally, act as a diagnostic aid for doctors (e.g. radiologists) if applied in real world and hosted on cloud (e.g. for individual doctors) or in medical entities on premise servers (e.g. clinics, or hospitals).

There are several machine learning models for processing different kind of problems based on data types. When processing images Convolutional Neural Networks (CNN) considered by far the most popular and widely used model, this special type of neural network consists of connected neurons and trainable parameters called weights that is particularly powerful in analysing image datasets.

3.6.1 Model Training – First Research Question

The proposed model is built for image classification using Chest x-ray dataset that contains images belongs to two classes (i.e. Pneumonia/Normal). Author used a subset of these images in a loop of 13 times in order to investigate the first research question (i.e. how many images is good enough to achieve relatively high performance results?). In this step, after loading images into TensorFlow dataset, author created image data generator for training set. Image augmentation been applied in the training set (e.g. zooming, flipping, and shifting) to prevent overfitting from happening in the network. When loading images, target size of images is set to (200x200) pixels, these resized images are loaded in 32 batches representing the two classes as two seeds and binary classification mode. After that, CNN is defined using Conv2D layer that is has many parameters to set (e.g. activation function set to ReLu). Dropout layer is applied using pooling filter of 2 by 2 to help in avoiding overfitting. Next, the network is run for compilation suing binary class and accuracy since it involves two classes. Finally, in order to minimize some loss function, the CNN model is trained by applying 10

epochs, this tend to improve accuracy at the end of looping a number of iterations. After setting up CNN architecture, author can put the model into work and start to make predictions by testing the model on non-seen images.

3.6.2 Model Training – Second Research Question

For the second research question, how does using (clean vs. dirty) and (balanced vs. unbalanced) training data impact models in getting more accurate outcomes? author proposed a solution based on Google Cloud AutoML Vision platform, an automated machine learning tool that allows to build a model that help doctors (e.g. radiologists) in investigating how can properties of data impact performance of models.

3.6.3 Model Training - Risks and Proposed Solution

Generally, when training data there are some common issues that may significantly affect model outcomes when training the targeted data. Firstly, the distribution of the data (i.e. unbalanced or biased data). This occurs when certain types of data are trained (e.g. learned) more than other types, this may tend to make bias towards these certain trained classes. Furthermore, using training data that are different from the real world one even when they fall into the same class, this would fail the model to classify accurately. In addition, mislabelling or dirty data in some complex images would also make it unclear for the model where any particular type should be classified. Therefore, it is important in some cases to do manual labelling by assigning these complex cases to expert medical stuff (e.g. radiologists). Besides that, having insufficient data for training can cause fuzziness and impact model performance, thus the amount of required data varies widely based on a number of factors (e.g. complexity of data, and datatype).

3.6.4 Model Training - Proposed Solution

To overcome this bias and prevent it from expansion, in this paper, while more training data will often lead to a better model, author will show what is the relatively acceptable performance results (i.e. above 90% accuracy) by adding images cumulatively to the proposed CNN model. Furthermore, author will show the importance of having a balanced, cleaned training data in sustaining quality-imaging data used for training the proposed model. On the other hand, author needed to use a google account to create a project in Google Cloud AutoML platform, define settings, and be able to use the associated google cloud storage for uploading dataset images, and to be able to get benefit of other google cloud resources and services. Using google services is paid and must have a valid credit card to charge the account. However, google gave a free balance (i.e. 300 dollars) as a free trail, so the author gets the benefit of this trail to solve the second research question and sub concerns issued from it.

3.7 Model Evaluation

There are many metrics that could be used for evaluating different DL models (i.e. Accuracy, Confusion matrix, F1 score, False positive rate, False negative rate, Log loss, Negative predictive value, Precession, Recall, ROC Curve, Specificity). In this paper, for the first research question, author most interested in finding (Accuracy) as an evaluation mechanism when applying the proposed model over different amount of data, that is because as any model (i.e. Normal - No Pneumonia) might be predicted most of the time. Nevertheless, what is important and author care about is a tool that rarely miss an actual Pneumonia case. This can be achieved more precisely by keep adjusting the number of classification number cumulatively to some extent. However, for the second research question, we realized that using (Precision and Recall) could be more effective for gaining better outcomes. Precision can answer our question (of all predications of a Pneumonia, how many were predicted right?)

and Recall can answer our question (of all real Pneumonia, how many did we detect?).

Furthermore, author chosen other important metrics list that can help in evaluating the proposed model according to the research question and most effectively (Precision, Recall, False positive rate, False negative rate).

3.8 Testing Hypothesis

In real world, once the model is trained then tested, and evaluated to measure its effectiveness to have satisfactory results, the model will be ready to make predictions in actual medical tasks using new unseen x-ray images.

The process of testing the proposed hypothesis and answering research questions through ML is an iterative problem solving technique and evolving one. This is important to tailor with the use case author trying to solve. This can be achieved by repeatedly reinvestigating the data quality (e.g. cleaning, balancing, and modifying some parameters) and evaluating metrics, that author has been chosen previously, to check model performance and efficiency.

The first research question implemented on premise hardware using software that runs python code along with its different library's on Anaconda platform. The second research question, been implemented on cloud using Google Cloud AutoML Vision platform.

3.9 Model Procedure and Tools

Based on the different research questions, Author used two procedures to implement the proposed model over the same dataset in order to leverage the use of open source technology for the benefit of getting more diversity results. The first one, is using laptop computer equipped with the following hardware specifications, Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz, 8 GB Installed RAM, NVIDIA GeForce 930MX GPU, 64-bit Windows 11 Pro operating system x64-based processor. In software the following tools were

used, Anaconda3 2023.09 Navigator 2.5.0, Python 3.11.5 | packaged by Anaconda. However, for the other part of this paper, author used a solution based on Google Cloud AutoML Vision platform, this platform helped the author in building a classification model that would serve as the backbone of research question and sub tasks under investigation. Google Cloud AutoML is a tool does not need any hardware or software to run (i.e. it does not need any code).

Chapter 4 : Experimental Results

In this section, author implemented the proposed methodology of this paper using experimental techniques to present how the model been used to evaluate the outcomes. Subset images of the chest X-Ray dataset was used to test models efficiency.

CNN model was used for implementation and present answers for the following research questions:

RQ1: How many images is good enough to achieve relatively high performance results?

RQ2: RQ2: How might the detection of pneumonia be impacted using chest x-ray images taken from different subsets of training data using different classifier models?

4.1 RQ1 Experimental Results

For answering the research questions of this paper, regarding the first main question:

RQ1: how many images is good enough to achieve relatively high performance results?

Author divided it into three sub research questions, as follows:

RQ1.1 how many images is good enough to achieve relatively high Categorical Accuracy percentage?

RQ1.2 how many images is good enough to achieve relatively Low Loss percentage?

RQ1.3 how many images is good enough to achieve relatively minimum Run Time in correlation with both Categorical Accuracy and Loss?

4.1.1 RQ1.1 Experimental Results

Figure (9) shows a line graph with a horizontal axis labeled (Number of Training Images) and the vertical axis labeled (Categorical Accuracy Percentage). The figure shows the curve of Categorical Accuracy in response to the number of images been implemented in

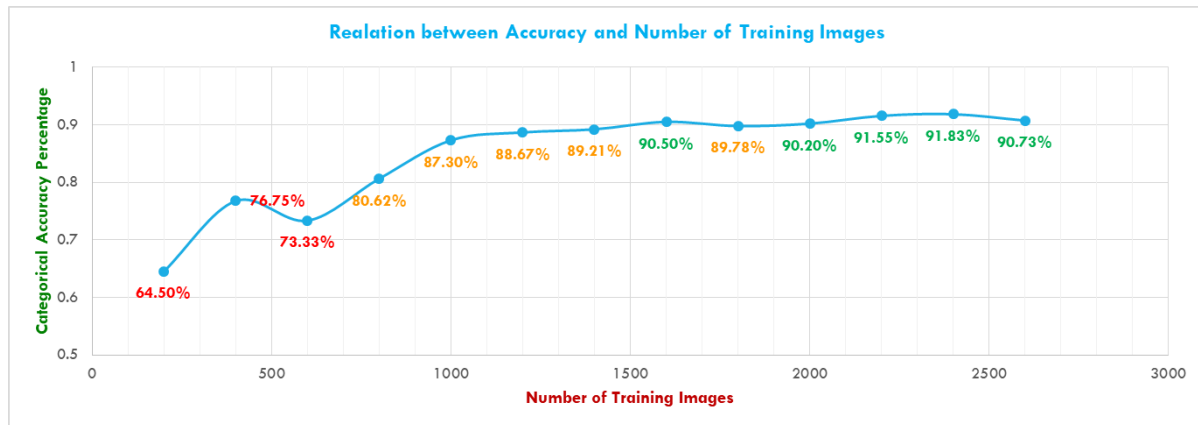
accordance to each run of the proposed model. Each data point represents the results in terms of percentage.

The graph illustrates how CA is rising up in correlation with number of training images.

Firstly, the percentage of CA rose a little starting at around 64.5% using only 200 images reaching 76.75% CA using 400 images, where CA depreciated to 73.33% using 600 images, then CA increased again to 80.62% using 800 images and grow gradually until 90.50% CA using 1600 images. After that, the line graph fluctuated up and down until the last data point where CA percentage is 90.73% using 2600 images. However, using 2400 images represented the greatest percentage of CA 91.83% as the highest data point level on the line graph.

Figure 9

Accuracy Curve



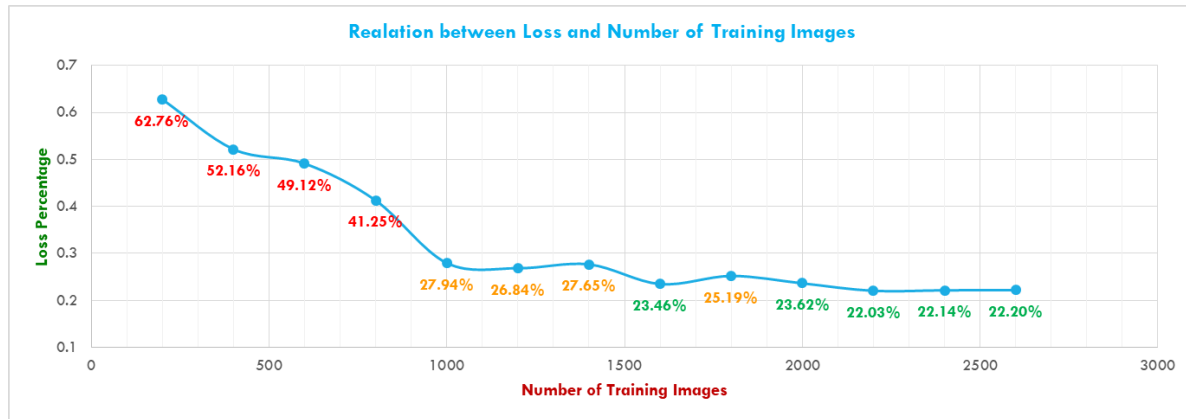
4.1.2 RQ1.2 Experimental Results

Figure (10) shows a line graph with a horizontal axis labeled (Number of Training Images) and the vertical axis labeled (Loss Percentage). The figure shows the curve of Loss in response to the number of images been implemented in accordance to each run of the proposed model. Each data point represents the results in terms of percentage.

The graph illustrates how Loss is decreasing in correlation with number of training images. Firstly, the percentage of Loss decreased gradually starting at around 62.76% using only 200 images until reaching 26.84% Loss using 1200 images, where Loss raised up a little bit to 27.46% using 1400 images. After that, the line graph slightly undulated up and down until the last data point where Loss percentage is 22.20% using 2600 images. However, using 2200 images represented the least percentage of Loss 22.03% as the lowest data point level on the line graph.

Figure 10

Loss Curve



4.1.3 RQ1.3 Experimental Results

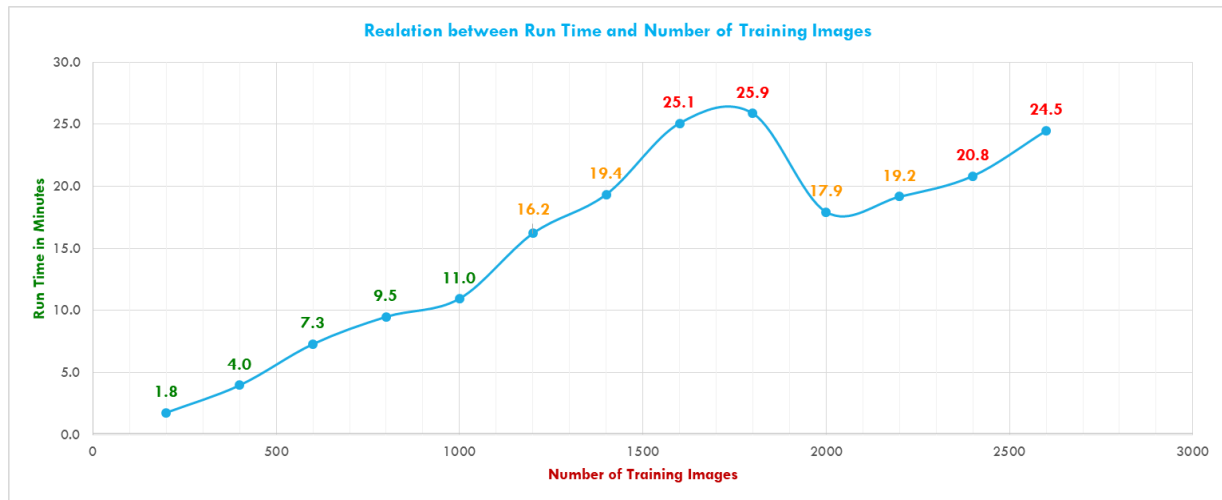
Figure (11) shows a line graph with a horizontal axis labeled (Number of Training Images) and the vertical axis labeled (Run Time). The figure shows the curve of Run Time needed in response to the number of images been implemented in accordance to each run of the proposed model. Each data point represents the results in terms of minutes.

The graph illustrates how Run Time RT duration is rising up in correlation with number of training images. Firstly, the duration of RT start rinsing at data point 1.8 minutes using only 200 images reaching the range of 25 minutes of RT using 1400 to 1600 images.

After that, RT declined to 17.9 minutes using 1800 images, then again, RT increased gradually growing up to 24.5 minutes using 2600 images. Using 1600 images represented the largest duration of RT 25.9 minutes as the highest data point level on the line graph.

Figure 11

Run Time Curve



4.2 RQ2 Experimental Results

In this paper, author created a model and hosted it on the cloud using Google AutoML, an automated ML tool. Then created classifying models using four different dataset variations to investigate how might the quality of training data affects the detection of pneumonia using the proposed classifying models.

For answering the research questions of this paper, regarding the second main question:

RQ2: How might the detection of pneumonia be impacted using chest x-ray images taken from different subsets of training data using different classifier models?

Author divided it into four sub research questions, as follows:

RQ2.1 How might the detection of pneumonia be accurate by using a binary classifier using chest x-ray images taken from balanced and clean data?

RQ2.2 How might the detection of pneumonia be accurate by using a binary classifier using chest x-ray images taken from unbalanced but clean data?

RQ2.3 How might the detection of pneumonia be accurate by using a binary classifier using chest x-ray images taken from balanced but dirty (unclean) data?

RQ2.4 How might the detection of pneumonia be accurate by using a three class model using chest x-ray images taken from balanced and clean data?

Author used four variations of the pneumonia dataset to train four distinct models. The dataset includes x-ray images of children's chests and have been divided into two categories (classes), normal and pneumonia. The steps author took to generate each model are outlined in the following sections.

4.2.1 RQ2.1 Experimental Results

RQ2.1 How might the detection of pneumonia be accurate by using a binary classifier using chest x-ray images taken from balanced and clean data?

To begin, author used 100 images of class Normal and another 100 images of class Pneumonia. That means 200 total chest x-ray images used to train the proposed model. In addition, author split these images into three sets training, validation, and testing where data distributed as 80%, 10%, and 10% respectively as shown in figure (12).

Figure 12*RQ2.1 Data Split*

Total images	200
Training images	160
Validation images	20
Test images	20

Author calculated Precision and Recall measurements. Setting up the data split as mentioned above (80% training, 10% validation, and 10% testing), and Confidence threshold to 0.5, resulted in Precision and Recall percentages equals to 100%, as shown in figure (13).

Figure 13*Precision and Recall Results in Percentage*

Confidence threshold	?		0.5
All labels			
Average precision	?	1	
Precision	?	100%	
Recall	?	100%	

4.2.2 RQ2.2 Experimental Results

RQ2.2 How might the **detection** of pneumonia be accurate by using a binary classifier using chest x-ray images taken from unbalanced but clean data?

In this part, author used the same count of images for of class Normal (100 images) but added 200 more images of class Pneumonia, that means the total count for pneumonia class equals to 300 images. In other words, this means 400 total chest x-ray images used to train the

proposed model. In addition, author split these images into three sets training, validation, and testing where data distributed as 80%, 10%, and 10% respectively as shown in figure (14).

Figure 14

RQ2.2 Data Split

Total images	400
Training images	320
Validation images	40
Test images	40

Setting up the data split as mentioned above (80% training, 10% validation, and 10% testing), and Confidence threshold to 0.5, resulted in 95% Precision and 95% Recall percentages, as shown in figure (15).

Figure 15

RQ2.2 Precision and Recall Results in Percentage

Confidence threshold		0.5
Average precision		0.973
Precision		95%
Recall		95%

4.2.3 RQ2.3 Experimental Results

RQ2.3 How might the detection of pneumonia be accurate by using a binary classifier using chest x-ray images taken from balanced but dirty (unclean) data?

In this part, author used 100 images of class Normal and another 100 images of class Pneumonia. That means 200 total chest x-ray images used to train the proposed model.

However, in this part, author switched the labels of 30 images in each class. That means 30% of the data are mislabelled as shown in table (2).

Table 2

Class Specific Data Distribution

Class	Clean images	Dirty images	Total
Normal	70 Normal Class	30 image from Pneumonia class	100 images
Pneumonia	70 Pneumonia class	30 images from Normal class	100 images

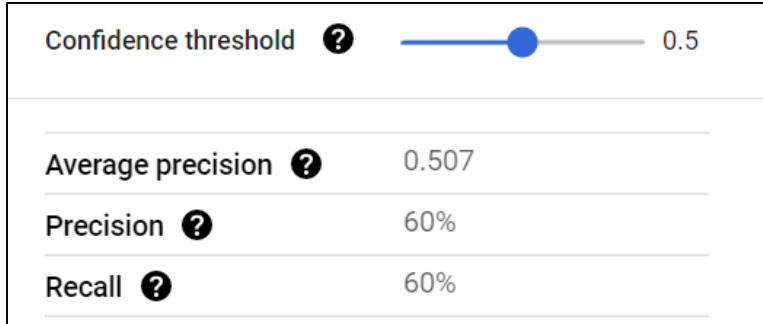
As in previous parts, author split these images into three sets training, validation, and testing where data distributed as 80%, 10%, and 10% respectively as shown in figure (16).

Figure 16

RQ2.3 Data Split

Total images	200
Training images	160
Validation images	20
Test images	20

Author calculated Precision and Recall measurements. Setting up the data split as mentioned above (80% training, 10% validation, and 10% testing), and Confidence threshold to 0.5, resulted in 60% Precision and 60% Recall percentages, as shown in figure (17).

Figure 17*RQ2.3 Precision and Recall Results in Percentage***4.2.4 RQ2.4 Experimental Results**

RQ2.4 How might the detection of pneumonia be accurate by using a three class model using chest x-ray images taken from a balanced and clean data?

For the final part, author get used of the benefit of having two different classes of Pneumonia (i.e. Bacterial Pneumonia and Viral Pneumonia). Thus, in order to implement the research question over a three class model, author used 100 images of class Normal and another 100 images of class Bacterial Pneumonia as well 100 images of class Viral Pneumonia. That means 300 total chest x-ray images used to train the proposed model. As in previous parts, author split these images into three sets training, validation, and testing where data distributed as 80%, 10%, and 10% respectively as shown in figure (18).

Figure 18*RQ2.3 Data Split*

Total images	300
Training images	240
Validation images	30
Test images	30

Setting up the data split as mentioned above (80% training, 10% validation, and 10% testing), and Confidence threshold to 0.5, resulted in 86.21% Precision and 93.33% Recall percentages, as shown in figure (19).

Figure 19*RQ2.4 Precision and Recall Results in Percentage.*

Confidence threshold ?	0.5
Average precision ?	0.85
Precision ?	86.21%
Recall ?	83.33%

Moreover, the precision and recall rates were calculated by the average precision and recall rates of all three classes.

From the above results, author may calculate one of the models' F1 score (e.g. RQ2.4) as follows:

The F1 score = $(2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$

= $(2 * 86.21\% * 83.33\%) / (86.21\% + 83.33\%) = 84.75\%$ so the F1 score = 0.8475

Chapter 5 : Summary and Discussion

5.1 RQ1 Results Summary

In Table (3), shows the efficiency of implementing the proposed model over certain amount of training images, a range between 200 and 2600 images, and differs in each run incrementally. More particularly, the table illustrates some interesting data about number of chest x-ray images used for training the proposed CNN model. It allows model clarification in outcomes when using different number of images as a total of both classes Pneumonia and Normal cases, and shows the percentage of Categorical Accuracy, percentage of loss, and Runtime of each round in minutes. Serial represents the number of times that the CNN model run. Loop of 10 epochs is set in each run. In the first run, author put 100 image from each class into training, which is 200 images in total to start with. In each run, the number of images was increased by 200 images in total for both classes, which means 100 image for each class. The table also shows that author run to model 13 times in order to able to find a fair ground to compare the results of each run. Thus, 2600 images were trained in final round.

In terms of Categorical Accuracy, in run number 12, shown the highest percentage of Accuracy equals to (91.55%) compared to the lowest percentage in the first run (64.5%). In other words, the difference between the highest and lowest categorical accuracy is (27.02%) where the total number of images is 12 times larger in run number 12 (2400 images) compared to only (200 images) used in the first run. In the final run of the model, where the number of total images was 2600, Categorical Accuracy achieved (90.73%) that is a bit lower than it is in the previous run by (0.82%).

In terms of Loss, in run number 11, shown the lowest percentage of loss equals to (22.03%) compared to the highest percentage in the first run (62.76%). In other words, the difference between the lowest and highest Loss is (40.73%) where the total number of images

in run number one (200 images) compared to (2200 images) used in run number 11. In the final run of the model, where the number of total images was 2600, Loss achieved (22.2%) that is a bit higher than it is in the run number 11 by (0.17%).

In terms of Run Time, in run number 9, shown the highest time consumed in minutes' equals to (25.9) compared to the lowest time consumed in the first run (1.8) minutes. In other words, the difference between the highest and lowest run time is (24.1) minutes where the total number of images is 9 times larger in run number 9 (1800 images) compared to only (200 images) used in the first run. In the final run of the model, where the number of total images was 2600, Run Time needed (24.5) minutes that is a bit higher than it is in ninth run by (1.4) minutes.

Finally, in terms of Accuracy Difference (AC), the highest difference is (12%) using 400 images at the second loop of running the model. after that, AC began decreasing gradually in the next early loops from 12% to 7% where 1000 images were used. After that, AC dropped to 1% using 1200 images. Then AC oscillated between -1% and 1% when using images higher than 1200 up to 2600.

Table 3*RQ1 Model Efficiency*

Ser	Number of	Number of	Tot	Categorical	Los	Ru	Accura
ial	Pneumonia Images	Normal Images	al	Accuracy	s	Ti	cy
							Differe
							me nce
1	100	100	20	0.645	0.62	-	
			0		76	1.8	
2	200	200	40	0.7675	0.52	4.0	12%
			0		16		
3	300	300	60	0.7333	0.49	7.3	-3%
			0		12		
4	400	400	80	0.8062	0.41	9.5	7%
			0		25		
5	500	500	10	0.873	0.27	11.	7%
			00		94	0	
6	600	600	12	0.8867	0.26	16.	1%
			00		84	2	
7	700	700	14	0.8921	0.27	19.	1%
			00		65	4	
8	800	800	16	0.905	0.23	25.	1%
			00		46	1	

9	900	900	1800	0.8978	0.25	25.	-1%
10	1000	1000	2000	0.902	0.23	17.	0%
11	1100	1100	2200	0.9155	0.22	19.	1%
12	1200	1200	2400	0.9183	0.22	20.	0%
13	1300	1300	2600	0.9073	0.22	24.	-1%

In this paper author showed that using 1600 x-ray images achieved 90.5% accuracy that is considered relatively high using state of the art CNN model. Where as in [22] authors had to use 2000 x-ray images to achieve 87.84% overall accuracy using VGG16 model, and 91.32% using InceptionV3 model, and 92.93% using EfficientNetB0 model.

5.2 RQ2 Results Summary

In Table (4), shows the efficiency of implementing Google AutoML model over certain amount of training images, a dataset ranges between 200 and 400 images, and differs in each run to show the results variations based on the quality of training data. More particularly, it shows the percentage of Precision and Recall of each dataset combination. Data split represents the number of images that been divided into training, validation, and testing data.

Table 4*Performance Summery*

Serial	Data Type		Data split	Results		
Research Question	Pneumonia	Normal	Total number of images	Precision	Recall	<i>F1</i>
RQ 2.1	100	100	200 (160 training, 20 validations, 20 testing)	100%	100%	100%
RQ 2.2	300	100	400 (320 training, 40 validations, 40 testing)	95%	95%	95%
RQ 2.3	100 (30 of them Normal)	100 (30 of them pneumonia)	200 (160 training, 20 validations, 20 testing)	60%	60%	60%

			300			
			(240			
RQ 2.4	200 (100 Bacteria, 100 Viral)	100	training, 30 validations, 30 testing)	86.21%	83.33%	84.75%

In this paper author showed that using clean and balanced x-ray images achieved as in RQ2.1 100% precision and accuracy and F1 score using Google Cloud AutoML model. Where as in [22] authors achieved 82% overall Precision and 82.33% overall Recall using VGG16 model, and 87.54% overall Precision and 89% overall Recall using InceptionV3 model, and 88.3% overall Precision and 90% overall Recall using EfficientNetB0 model. Moreover, author found F1 score for RQ.2.4 (i.e. a balanced and clean classification model) and it equals to 84.75%, which close to the overall F1 score using VGG16 model.

Chapter 6 : Conclusion

There is no doubt that the Pneumonia infection poses a threat to human health especially among children. Efforts to control the infection's spread have been observed to place a strain on the healthcare system. The detection of the presence by testing methods is expensive and might not be sufficient to reach a larger population. In this paper, the author investigated two research questions that aims to enhance Image recognition and classification using ML methodologies. The data input of X-ray images taken for medical proposes, used to gain better outcomes through advanced analysis of the training data, this includes specifying the average amount of data needed for training to make a good enough predictions using deep learning (DL) in order to save costs. In addition, exploring training data by applying data cleaning techniques to gain a well-balanced model for classification purposes. Author shown that setting 1600 x-ray images or more, as a training data input, tend to enforce a steady percentage of accuracy greater than 90% since the Accuracy difference was less than 2% as average to Accuracy difference in all 13 loops and this will influence the classification process appeal to be worthless and much costly. In general, author noted that Accuracy runs in favour correlation with Loss Percentage and Run time results. However, having done the on laptop with traditional specifications makes doubtful results especially Run Time. Moreover, author described the results of using dirty (unclean) or unbalanced data to the ML model, which showed a clearly drop in precision, recall and F1 score percentages where author shown that quality training data has a great impact over these results (precision and recall percentages), thus maintaining a well-balanced and clean data contribute better outcomes. However, implementing the model on cloud is a paid service, and author used low count number of total images, this can be unsafe with limited resources. Overall, our proposed experiments showed the importance of having a quality training data in achieving higher performance results.

Chapter 7 : Future Work

Author will look more deeply and priorities the most affected cases in medical field, over larger set of data, over more historical and recent period of time. In addition, authors will dig more into deep learning models using CNNs in order to enhance its performance. Moreover, trying another types of x-ray images, and another types of images other than medical field to test the proposed model on another domain.

References

- Apostolopoulos, I. D., Aznaouridis, S. I. & Tzani, M. A. (2020). Extracting Possibly Representative COVID-19 Biomarkers from X-ray Images with Deep Learning Approach and Image Data Related to Pulmonary Diseases. *Journal of Medical and Biological Engineering*. Springer Berlin Heidelberg, vol. 40(3), pp. 462–469.
- Basu, K., Sinha, R., Ong, A. & Basu, T. (2020). Artificial intelligence: How is it changing medical sciences and its future? *Indian Journal of Dermatology*, vol. 65(5), pp. 365–370.
- Cao, Z., Duan, L., Yang, G., Yue, T. & Chen, Q. (2019). An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC Medical Imaging*. BMC Medical Imaging, vol. 19(1), pp. 1–9.
- Chang, J., Yu, J., Han, T., Chang, H. J. & Park, E. (2017). A method for classifying medical images using transfer learning: A pilot study on histopathology of breast cancer. *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom 2017*, vol. 2017-December, pp. 1–4.
- Cho, J., Lee, K., Shin, E., Choy, G. & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? [online]. Available at: <http://arxiv.org/abs/1511.06348>.
- Fisher, I. & Ziviani, J. (2004). Explanatory case studies: Implications and applications for clinical research. *Australian Occupational Therapy Journal*, vol. 51(4), pp. 185–191.
- Gao, F., Yoon, H., Wu, T. & Chu, X. (2020). A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications*. Elsevier Ltd, vol. 143, p. 112957.

- Gaur, L., Bhatia, U., Jhanjhi, N. Z., Muhammad, G. & Masud, M. (2023). Medical image-based detection of COVID-19 using Deep Convolution Neural Networks. *Multimedia Systems*. Springer Berlin Heidelberg, vol. 29(3), pp. 1729–1738.
- Greenstone, M., Harris, M., Li, K., Looney, A. & Patashnik, J. (2012). A Dozen Economic Facts About K-12 Education, (August), p. 24.
- Hebbale, S., Marndi, A., Achyutha, P. N., Manjula, G., Mohan, B. R. & Jagadeesh, B. N. (2022). Automated medical image classification using deep learning. *International journal of health sciences*, (August), pp. 1650–1667.
- Jeyaraj, P. R. & Samuel Nadar, E. R. (2019). Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *Journal of Cancer Research and Clinical Oncology*. Springer Berlin Heidelberg, vol. 145(4), pp. 829–837.
- Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V. A. N., Wen, C., Zhang, E. D., Zhang, C. L., Li, O., Wang, X., Singer, M. A., Sun, X., Xu, J., Tafreshi, A., Lewis, M. A., Xia, H. & Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*. Elsevier Inc., vol. 172(5), pp. 1122-1131.e9.
- Khan, S. & Yong, S. P. (2018). A deep learning architecture for classifying medical images of anatomy object. *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, vol. 2018-February(December), pp. 1661–1668.

- Kowsari, K., Sali, R., Ehsan, L., Adorno, W., Ali, A., Moore, S., Amadi, B., Kelly, P., Syed, S. & Brown, D. (2020). HMIC: Hierarchical medical image classification, a deep learning approach. *Information (Switzerland)*, vol. 11(6), pp. 1–19.
- Lai, Z. & Deng, H. (2018). Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron. *Computational Intelligence and Neuroscience*, vol. 2018.
- Liu, X., He, J., Song, L., Liu, S. & Srivastava, G. (2021). Medical image classification based on an adaptive size deep learning model. *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 17(3s).
- Mijwil, M. M. & Al-Zubaidi, E. A. (2021). Medical image classification for coronavirus disease (covid-19) using convolutional neural networks. *Iraqi Journal of Science*, vol. 62(8), pp. 2740–2747.
- Muheidat, F. & Tawalbeh, L. (2021). *Artificial Intelligence and Blockchain for Cybersecurity Applications*.
- Narayan, V., Mall, P. K., Alkhayyat, A., Abhishek, K., Kumar, S. & Pandey, P. (2023). Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images. *Journal of Sensors*, vol. 2023.
- Pneumonia in children*. (2022) [online]. Available at: <https://www.who.int/news-room/fact-sheets/detail/pneumonia>.
- Raj, R. J. S., Shobana, S. J., Pustokhina, I. V., Pustokhin, D. A., Gupta, D. & Shankar, K. (2020). Optimal feature selection-based medical image classification using deep learning model in intern. *IEEE Access*, vol. 8, pp. 58006–58017.

- Reinsel, D., Gantz, J. & Rydning, J. (2018). The Digitization of the World - From Edge to Core. *Framingham: International Data Corporation*, (November), p. US44413318 [online]. Available at: <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>.
- Roth, H. R., Lee, C. T., Shin, H. C., Seff, A., Kim, L., Yao, J., Lu, L. & Summers, R. M. (2015). Anatomy-specific classification of medical images using deep convolutional nets. *Proceedings - International Symposium on Biomedical Imaging*, vol. 2015-July, pp. 101–104.
- Stephen, O., Sain, M., Maduh, U. J. & Jeong, D. U. (2019). An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. *Journal of Healthcare Engineering*, vol. 2019.
- Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S. & Lu, J. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*. Elsevier Ltd, vol. 75, pp. 34–46.
- Wang, D. & Wang, L. (2019). On OCT image classification via deep learning. *IEEE Photonics Journal*. IEEE, vol. 11(5), pp. 1–14.
- Wang, G., Li, W., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S. & Vercauteren, T. (2018). Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. *IEEE Transactions on Medical Imaging*, vol. 37(7), pp. 1562–1573.
- Zhang, J., Xia, Y., Wu, Q. & Xie, Y. (2017). Classification of Medical Images and Illustrations in the Biomedical Literature Using Synergic Deep Learning [online]. Available at: <http://arxiv.org/abs/1706.09092>.

Zhang, J., Xie, Y., Wu, Q. & Xia, Y. (2019). Medical image classification using synergic deep learning. *Medical Image Analysis*, vol. 54, pp. 10–19.