Detecting and Understanding COVID-19 Misclassifications: A Deep Learning and Explainable AI Approach

Nkcubeko Umzubongile Siphamandla Mandindi

Supervisor(s):
Professor Hima Vadapalli

A research report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg

27 August 2023

Declaration

I, Nkcubeko Umzubongile Siphamandla Mandindi, declare that this research report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Nkcubeko Umzubongile Siphamandla Mandindi
27 August 2023

Abstract

Interstitial Lung Disease (IDL) is a catch-all term for over 200 chronic lung diseases. These diseases are distinguished by lung tissue inflammation (Pulmonary fibrosis). They are histologically heterogeneous diseases with inconsistent microscopic appearances, but they have clinical manifestations similar to other lung disorders. The similarities in symptoms of these diseases make differential diagnosis difficult and may lead to COVID-19 misdiagnosis with various types of IDLs. Because the turnaround time is shorter and more sensitive for diagnosis, imaging technology has been mentioned as a critical detection method in combating the prevalence of COVID-19. The aim of this research is to investigate existing deep learning architectures for the aforementioned task, as well as incorporate evaluation modules to determine where and why misclassification occurred. In this study, three widely used deep learning architectures, ResNet-50, VGG-19, and CoroNet, were evaluated for detecting COVID-19 from other IDLs (bacterial pneumonia, normal (healthy), viral pneumonia, and tuberculosis). The baseline results demonstrate the effectivities of Coronet having a classification performance of 84.02% for accuracy, specificity of 89.87%, a sensitivity of 70.97%, Recall 84.12%, and F1 score of 0.84. The results further emphasize the effectiveness of transfer learning using pre-trained domain-specific architectures, resulting in fewer learnable parameters. The proposed work used Integrated Gradients (IG), an Explainable AI technique that uses saliency maps to observe pixel feature importances, to understand misclassifications. This refers to visually prominent features in input images that were used by the model to make predictions. As a result, the proposed work envisions future research directions for improved classification through misclassification understanding.

Keywords: Coronavirus, Tuberculosis, Bacterial Pneumonia and Viral Pneumonia, Chest X-rays Radiographs, Convolutional Neural Networks, Transfer Leraning, Explainable AI, Scilency, Integrated gradience

Acknowledgements

I want to express my gratitude to my supervisor, Professor Hima Vadapalli, for providing me with the opportunity to work on this excellent research project titled "Detecting and Understanding COVID-19 Misclassifications: A Deep Learning and Explainable AI Approach," which has pushed me to perform a lot of research. Second, I'd want to express my gratitude to my sponsor, the National e-Science Postgraduate Teaching and Training Platform (NEPTTP), for making this opportunity a reality.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etworks
CNN	**C**rtificial **N**eural **N**etworks
COVID-19	**CO**rona **VI**rus **D**isease 19
CT	**C**omputed **T**omography
DL	**D**eep **L**earning
DNN	**D**eep **N**eural **N**etworks
ICTV	**I**nternational **C**ommittee on **D**isorders of **V**irus
IDL	**I**nterstitial **L**ung **D**isorders
IG	**I**ntegrated **I**radients
ML	**M**achine **L**earning
PPE	**P**ersonal **P**rotective **E**quipment
RT-PCR	**R**everse **T**ranscriptase **P**olymerase **C**hain **R**eaction
RNN	**R**ecurrent **N**eural **N**etworks
SARS-CoV	**S**evere **A**cute **R**espiratory **S**yndrome **C**orona**V**irus
SARS-CoV-2	**S**evere **A**cute **R**espiratory **S**yndrome **C**orona**V**irus 2
TB	**T**uberculosis
TL	**T**ransfer **L**earning
WHO	**W**orld **H**ealth **O**rganization
XAI	E**X**plainable **A**rtificial **I**ntelligence

Chapter 1

Introduction

Interstitial Lung Disease (IDL) is an umbrella term for more than 200 various types of chronic lung infections. The pathophysiology of these disorders is characterized by inflammation of the lung tissue (Pulmonary Fibrosis). Fibrosis causes the stiffening of the lung tissue which intern reduces the process of diffusion in the lungs which is the air sacs' capacity to transport and supply oxygen. IDLs are heterogeneous histologically, meaning their microscopic appearances are somewhat inconsistent but have similar clinical manifestations to other lung disorders, also seen in novel coronavirus disease for 2019 (COVID-19) cases [48]. In most cases, the similarities in symptoms of these diseases make it difficult to determine the differential diagnosis and may lead to misdiagnosis of COVID-19 to other profuse forms of bacterial and viral pneumonia and tuberculosis (TB) [3]. The gold standard for identifying nucleic acids of a virus has been real-time reverse transcriptase polymerase chain reaction (RT-PCR) [28]. Unfortunately, due to the high number of false-negatives results, the waiting period to get test results, and the shortage of medical supplies has made the fight against COVID-19 a strenuous and unsafe task [37][51][50][17]. This shortage includes the supply of RT-PCR assay kits[48][28], ventilators, and personal protective equipment (PPE)[37] that has affected many countries due to increased demand. This has made health practitioners and scientists look for a better and faster alternative for diagnosis: the use of imaging technologies like an X-ray or computed tomography (CT) scans [3][16][28].

Imaging technology has been mentioned as being a critical detection method in fighting COVID-19 as it is easily performed when a patient comes into a healthcare facility and the result turnaround time is highly quick and sensitive for diagnosis. This can result in immediate treatment and care, as well as isolation if the patient is deemed positive [13][65][50]. It is also suggested that this method reveals the amount of damage, as shown by hazy lung opacities such as air space consolidation, as well as the probable pathology's precise location [39][48]. However, the accuracy and precision at which these images have been classified as a constraint in adopting the technology. Several countries have experienced and continue to experience uncertainty due to COVID-19's effects on the economy and health. It has been difficult for front-line healthcare personnel without radiology training to use X-ray pictures correctly and effectively[51]. Previously, it was discovered that potential Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2) infection indications might exist in uninfected SARS-CoV-2 people, and the differences in their manifestation may be rather minute[17].

Using chest X-rays, this study tries to determine the best-performing image classification model that distinguishes COVID-19 from other common IDLs. In this experiment, methods to optimize the performance and accuracy of the model's prediction ability will be investigated. Here, we seek to identify where and why misclassifications occur within these models. We hypothesize that if we get the best image classification model and can understand where misclassification occurs to improve the models' performance, we will be able to accurately diagnose patients that come into a healthcare facility and have a quick turnaround time that is sensitive for diagnosis. This is to reduce exposing non-infected patients and healthcare workers to infected individuals.

1.1 Background

In this section, we will review topics that help justify the current research in detail. We will look into one of the most popular and recent IDLs that has caused the world to go to a halt and economic distress, COVID-19. This is to show how medical imaging can help combat the spread of this deadly disease. Section 1.1.2 will describe what deep learning is, where it has been applied, and justify why it will be adopted in this research. Lastly, Section 1.1.3 will look into misclassification and Explainable AI to help us understand why images may have been misclassified.

1.1.1 Novel Coronavirus-19

On February 11, 2020, the World Health Organization (WHO) announced a transmissible illness caused by SARS-CoV-2 as the Novel Coronavirus Disease for 2019. (COVID-19) [16][48][57]. The International Committee on Taxonomy of Viruses (ICTV) guidelines were followed for this designation[56]. First discovered in December 2019 in Wuhan, China, COVID-19 rapidly spread through the borders of most countries around the world [16][13][65]. According to the John Hopkins Resource Centre, to date, COVID-19 has global cases amounting to 754, 018, 841 and a death toll of 6, 817, 478 people [19][30]. The virus was initially contracted at the Wuhan Market in the Province of Hubei, China, by tourists that were visiting. It is, therefore, believed but not confirmed that the disease would be of animal origin (zoonosis)[16].

Similar viruses, such as Severe Acute Respiratory Syndrome (SARS-CoV) in 2003 and Middle East Respiratory Syndrome (MERS-CoV) in 2012, would last for weeks or even months in the previous two decades [3][61]. COVID-19 had been standing for three years and counting at the time of writing this research report. COVID-19 has been announced as the World Health Organization's sixth public health emergency of global significance this decade [48]. The gradual onset of symptoms, which led to extensive transmission by asymptomatic carriers, is one reason underlying the outbreak of occurrences[48]. The symptoms of COVID-19 include fever (38 C or higher), dry cough, and dyspnea [39][16][54]. Other sources have included symptoms like an initial throat infection[61], tiredness[3], the loss of smell without an obstruction in the nose(Anosmia), and a total loss in taste(Ageusia)[16].

The mode of transmission of this disease, droplets, makes it extremely lethal. These are secretions excreted from infected individuals' nasopharyngeal tracts through modes of talking, coughing, and sneezing [16]. Thus, being in close proximity to infected individuals and touching surfaces where the spread is through contact with fomites once an infected individual covers their mouth while sneezing and coughing. While vaccines are being developed and distributed, WHO and governments in many countries around the world proposed and implemented preventative measures. These precautions include hand sanitation with a hydroalcoholic solution of $\geq$70 percent alcohol, avoiding close contact employing social distancing $\geq$1 meters from one another, avoiding shaking hands, avoiding touching your mouth, nose, and eyes, and most importantly always wearing a protective musk of the nasopharyngeal region[16].

1.1.2 Deep Learning

Deep learning is not the same as traditional machine learning. It allows computational models with multiple neural network-based processing layers to assimilate data representation at different levels of abstraction [75]. Furthermore, there are distinctions between deep learning and traditional artificial neural networks (ANN). ANNs are only taught to obtain supervised representations for the optimization of specific objectives, whereas deep learning algorithms create observable pattern representations from input data obtained from the layer before or by optimizing unsupervised metrics [62]. Deep learning has had a tremendous impact on a plathora of applications, namely Speech Recognition [1], Computer Vision [45], and Natural Language Processing [43].

Deep learning has provided potential solutions to a plethora of computer vision-based healthcare problems [63]. Deep learning architectures applied in the medical imaging domain range across the medical sphere. Prasoon et al.[12] used Convolutional Neural Networks (CNN) to leverage knee MRIs to predict the risk of osteoarthritis. The 114 unseen scans were utilized to train the classification integrating three 2D Convolutional Neural Networks (CNNs). The experiment yielded 99.91% pertaining to accuracy, 78.58% for sensitivity, and specificity with 99.95%. Esteva et al.[2] used a collection of data of 129,450 clinical photographs of pixels and annotated images to train a CNN to identify skin lesions that might have been malignant. With the help of two crucial binary classifications and biopsy-proven clinical photos, the model's performance was compared to that of 21 board-certified dermatologists. In both tests, the CNN outperformed all tested professionals, proving that deep learning has the potential to classify skin cancer with a degree of proficiency on par with dermatologists. Employing a deep learning-trained system, an algorithm was created in retinal fundus pictures for the automated detection of diabetic macular edema and diabetic retinopathy. The model had a Receiver Operating Curve (ROC) area of 99.1 % and a sensitivity of 87.0% after being trained on a retrospective development data set containing retinal images amounting to 128,175 [73].

The use of imaging technology (X-ray or CT)) as a diagnosis source for COVID-19 has perceived higher sensitivity and thus, suggested to be an alternative to RT-PCR assay [28]. Radiologists and experts typically interpret X-ray images based on experience, and they frequently have different interpretations of the same image. This is due to the large amount of interpretation that they must perform, and fatigue usually leads to a misdiagnosis. As a result, computer-aided diagnosis systems are being used to assist radiologists and experts in interpreting images faster and more accurately [3][48][28].

A CNN is an effective deep learning algorithm in computer vision[63] and has been adopted to solve a variety of problems for image processing in the medical field [12][73][2]. However, new CNN architectures have been suggested for COVID-19 detection, namely CoroNet [**ref26**][66] , nCOVnet [38], COVID-Net [10][51][17][50][37], CovidxNet [31] , DarkCovidNet [71]. Deep learning architectures are data-hungry and necessitate a large amount of data for network training. This requirement can be mitigated to some extent by using Transfer Learning (TL), which is widely used to transfer and re-use knowledge across domains when data samples for a task are limited. By adjusting the final fully connected layers, pre-trained CNNs such as ResNet-50, VGG, and DenseNet-201 can be used to learn new tasks [3]. TL is said to favor the hypothesis that training data should be identically distributed and independent of the test data [13]. These pre-trained networks were developed using the ImageNet dataset [58] where images of 1, 000 classes are continuously updated. Transfer learning has previously been applied in several studies to identify COVID-19 from XR scans [**ref26**][66]. Deep Feature (DF) extraction does not necessitate network retraining. DF extraction directly extracts features from the deep learning network's final completely connected layer. These would be the CNNs listed below ResNet-50, VGG-19, and AlexNet are three networks [3]. The difference between the two methods is the computational time. Transfer learning requires more training time despite its simplicity in application. Yet, with DF extraction, it is the saving of computational time and resources that deems it advantageous to use [3][48].

1.1.3 Misclassifications and Explainable AI

Because of their complexity, deep learning models are known as black-box approaches. As a result, humans do not have a complete understanding of what is going on inside them [60][6]. DNNs are the most well-known black-box algorithms, yet, because of their ability to dominate in performance, they are extremely popular [6]. Deep Learning classifiers are used in the real world for a variety of tasks, but they frequently fall short when the sample distributions used for training and testing are different. Unnoticed failures occur when a high prediction confidence masks incorrect predictions and classifiers fail to indicate when and why they are incorrect. This could have negative consequences in a variety of areas. In medical diagnosis, for example, models that classify incorrectly with high confidence may result in incorrect decisions and instill an untrustworthy machine learning technology [26][6]. A classifier's goal is to offer the robust trustworthy prediction and should stand even when exposed to noisy samples. The existing architectures face uncertainty and ideally be able to concede their ignorance and misclassifications [18]. It is thus important to identify and develop tools or methods that improve the reliability and performance of DL architectures.

Systems that make decisions that cannot be well-interpreted are not trustworthy. This is of significance in sectors where human lives may be at risk. These include: healthcare [60][6] or self-driving cars [44]. The need for trustworthy models, models that are fair, robust, and perform well in real-world applications, has resurrected the field of Explainable AI. It is a branch of research concerned with understanding and interpreting the behaviors of AI systems [60][29], in particular, how they got to a prediction or decision [6]. Explainable AI has attempted to bridge the gap between the robust performance of ML models and the ability to foster explainability, interpretability, and causality [60][5][6]. First, let's define the terms mentioned above. Explainability is the ability to explain why a decision was made by observing the model's behavior. Technically, it should highlight the characteristics that contribute to the classification decision and attempt to explain why they are important at what magnitude. Causality is the ability to map human intelligence and the level of understanding humans have with AI explanations [5]. The ability to explain and present findings in terms that humans can understand is referred to as interpretability. In this case, knowing the weights and features allows us to easily calculate the output and answer precisely how, and with what functional mechanics, the model arrived at the predicted decision. Explainability and this phrase are frequently used interchangeably. All of this is connected to the ML approaches that must be explained. Machine learning approaches are classified into two types. Black-box models include deep learning and ensemble models. Models with obvious outcomes, such as white-box or glass-box models, exist. Linear models and tree-based models are examples of glass-box models [60]. In the following paragraphs, we will attempt to list and explain these XAI methods in order to be able to explain the decisions these black-box models make.

Several XAI strategies have been developed to help bridge this knowledge gap by allowing human experts to identify and comprehend the explanatory factors underlying an AI prediction. We will discuss popular methods that have influenced the development and growth of this application. The last decade has contributed to this growth, as XAI method developers have created model-agnostic and model-specific methods that can take either local or global explanations. A local method is one that explains only one specific instance. A global method is one that explains the entire model [60][6]. These methods are classified based on their explanation goal. These methods are functional, archetypal, and ad hoc. When it comes to archetypal methods, they try to find an input pattern x that is an original epitome of y. Consider activation maximization, in which the model seeks a randomized input pattern that will result in the best possible reaction or signal from the model. Models developed include generative adversarial network models. [33]. Functional approaches, on the other hand, inspect the learned representation in the graph to reveal algorithmic mechanisms. The methods used in this approach seek to illuminate how feature maps learn layer by layer in order to support their overall explanation. Finally, post-hoc methods try to understand the aspect of input x that makes it more likely to choose a class. These methods are classified into three types. Examples include feature ranking, perturbation methods, and decomposition methods. Methods for feature ranking and selection have been around since the early 2000s [41]. Perturbation relevance methods provide local estimates of an image's important regions or features. These can be model agnostic, depending on how the model is implemented. Decomposition methods, also known as redistribution methods, examine the model's structure and attempt to identify the share of relevance through the model layers. These attribute methods return the message and suggest relevance to the model, and then view predictions as a computational graph output [33].

The majority of these are classified as post hoc methods because they attempt to explain the models. Layer-wise Relevance Propagation (LRP), which is deduced on Deep Taylor Decomposition, and Local Interpretable Medels Agnostic Explanations (LIME), which uses an interpretable attribute space and local approximations with sparse K-LASSO, were among the first to explain these complex back-box model decisions [6]. The gradient explanation technique is a DNN interpretability method. It is a gradient-based attribution method that evaluates how much each input dimension changes the prediction in a relatively tiny area grouped around the input for each gradient [46]. Another technique is Integrated Gradients (IG), which explains DNN predictions by attributing to network input characteristics in its gradient-based attribution. It combines gradients between the baseline and the observation of interest. DeepLIFT, a different technique, was developed to supplement DNN predictions. This method improves on the gradient method's results by multiplying the gradient by the input signal, resulting in a first-order Taylor approximation. When applied to natural photos and genomics data, this technique outperformed gradient-based techniques [15][60]. Both IG and DeepLIFT have been used in medical tasks such as protein and DNA binding sites, and they were able to identify features associated with models with predictions of these for a sample in comparison to a baseline [60][8]. LRP, IG, and DeepLIFT are examples of decomposition methods because they seek to identify important features [33]. SHarperly Additive exPlainations (SHAP) is a model-agnostic, post hoc method that focuses on both understanding local interpretations and understanding global interpretations. It employs game theory to sharply value and unites the previously mentioned LIME, DeepLIFT, and LRP [6]. Testing with Concept Activation Vectors (TCAV) is a gradient-based strategy that is believed to leverage concepts acquired from annotated samples. [6]. SHAP and LIME are methods that are model agnostic and belong to the perturbation class [33]

Explainability is a widely discussed and emphasized topic in medicine. From the perspective of health stakeholders, it is not so much about performance as it is about why an AI-driven system makes a particular decision. The emphasis has been on comprehending the model's characteristics that influence decision-making [5][29][6]. The ability to divide interpretability or explainable methods is proportional to the scale of interpretation. To put it another way, interpretability is the prerequisite for other important characteristics such as privacy, dependability, fairness, trust, usability, transparency, and causality to be embedded with intelligent medical systems [33]. A brief summary of the methods are as follows.

Layer-wise Relevance Propagation (LRP) is an explanation technique that falls under the propagation class that requires weights and activations. With this information, LRP can solve the explanation problem by simplifying it. LRP exposes and employs the neural network structure, redistributes the explanatory factor layer by layer, and this is known as relevance (R), starting with the outcome of the model and working your way down to the source variables (e.g., pixels). LRP's central notion is to redistribute the whole result of the outcome $f(x)$ iteratively using decomposition [6]. This is explained in the following concept: $\sum_i R_i^{(0)} = ... = \sum_j R_j^{(l)} = \sum_k R_k^{(l+1)} = ... = f(x)$,
Where $R_i^{(0)}$ denotes the relevence assigned tot the ith input element, which could be the input, while $R_j^{(l)}$ stands for the nearon j ath the lth layer. The conservation redistribution assures that no significance is included or removed along the route, and it also allows for signed explanations, with positive significance values representing evidence confirming the forecast and negative relevance values representing evidence refuting it [6].

Local interpretable Model-agnostic Explanations (LIME), also known as the Surrogate-based explanation, seeks to clarify a prediction of a complicated system f_m, such as deep neural networks, by matching a proxy model f_s with simple outputs. LIME essentially produces sample data in the N_{x_i} neighborhood of the input we are interested in x_i [60]. The target model is then used to evaluate these, and the target model is then approximated in this local space using a simple linear function. LIME does not explicitly describe the target algorithms forecasts $f_m(x_i)$, but rather the projections of the surrogate model $f_s(x_i)$ [6]. This is represented by the following: $f_M(x) \approx f_S(x) for x \in N_{x_i}$.

Shapely Additive exPlanations (SHAP) is a model-independent, post-hoc local interpretability method that can be applied to any data set to obtain a global understanding. This algorithm explains individual predictions made by machine Learning models using Shapley values. It is concerned with elucidating a single point x^* for the model f, in which it is predicated on a functional form e_S, where S is a variable index subset $S \subseteq \{1, ..., p\}$. This component is described as the anticipated value given a prior distribution with this qualification applied to all factors in the subset S. This equation represents this concept $e_S = E[f(x)|x_S = x_S^*]$ [60]. This function is defined for extra data modalities as what the model anticipates at x^* after zeroing out the values for parameters with indices greater than one in S. The value e_S represents the model's response to the specification of variables within subset S. This approach uses Sharpley Additive Values to allocate the unequal contributions of all team members and is based on the theoretical underpinnings of game theory to explain the output of any ML model. The method follows an axiom in which features which do not make a contribution to the prediction are assigned a value of zero. To interpret the model globally, shapely values can be consolidated, and feature reliance plots, feature importance, and interaction analysis can be used [6]. Sharply values, on the other hand, are computationally complex, particularly with deep neural networks, where high dimensional inputs are unmanageable. Certain Shap estimation versions, such as Kernal SHAP (LIME with Shapley Smoothing Kernal), Tree SHAP (Shapley estimates for Trees), and Deep SHAP (Shapley estimates for Neural Networks), can be used [60].

In the background section, we sought to explain the three topics that the study will revolve around. The case study topic is COVID-19 and its rise to fame, Deep learning (DL), and its building blocks, and lastly, Explainable AI and its popular methods which include SHAP, and LRP just to mention a few. The next subsections seek to clearly mention the research question and research objectives and he clear rationale of the study.

1.2 Problem Statement

The use of RT-PCR in detecting viral nucleic acids has been the golden standard in diagnosis. Unfortunately, due to the high number of false-negative results, their adoption can cause undesirable outcomes. As a result, classifiers that fail to signal when they are likely to be incorrect can limit the potential to reduce COVID-19 spread. This has made health practitioners and scientists look for a better and faster alternative for diagnosis to prevent the outbreak of novel COVID-19.

1.3 Hypothesis

We hypothesize that the state-of-the-art CNN architecture pre-trained on COVID-19 images will outperform shallow and deep neural networks and thus will be able to help in the diagnosis of patients. Alternatively, shallow or deep neural networks will perform well compared to the other architectures. We also hypothesize that Explainable AI techniques will be able to help identify where the misclassification might have occurred in images classified in a similar class but are labeled as different classes thus eliminating the black box approach Deep Learning is associated with. Alternatively, Explainable AI will not be able to help us understand where these misclassifications occurred.

1.4 Research Question

1) Which classification models perform better between shallow networks, deep networks, and networks pre-trained on COVID-19 images?
2) When and where do classification models misclassify?
3) How can existing explainable AI techniques be able to explain why and where the misclassification occurs?

1.5 Research Aims and Objectives

This section states the rationale of this study in detail, by formulating and presenting a set of research aims and objectives. .

1.5.1 Research Aim

The study aims to identify which CNN architecture between a shallow network, a deep network, and a pre-trained state-of-the-art network will be able to detect a class of IDLs and to explain why misclassification occurred using an explainable AI technique.

1.5.2 Objectives

The objective of this study includes:

1) To train, validate and test the different CNN architectures on our dataset
2) To determine which CNN architecture will best predict IDLs and to evaluate the models' performance across COVID-19, normal, pneumonia caused by viruses, tuberculosis, and pneumonia caused by bacteria data.
3) To observe and compare how the performance improves between the original dataset and where data augmentation is applied.
4) To observe and understand where the misclassification occurred when Integrated Gradience (IG) is implemented for explainability.

1.6 Limitations

The key problem was the dataset's small sample size and the restricted representation of each class in the sample set available as open-source samples.

1.7 Conclusion

Finding the best-performing classification CNN architecture proposed for medical image detection, in this case being able to diagnose COVID-19 using chest X-rays, may contribute to the early detection of this deadly virus and help distinguish between the types of lung diseases. Using Explainable AI will help in paving a way for us to better understand why misclassification may occur and help build even more robust models for medical imaging. The flow of this research report is organized as follows: The methods that have been used will be detailed in chapter 3 for the purpose of giving more reference and understanding. Feel free to skip this section if familiar with the concepts. The methodology chapter will proceed with the methods chapter as the 4th chapter as it will describe and discuss the dataset and techniques adopted in this research report. The results will then be discussed in chapter 5 and we will then present a discussion section as chapter 6. The conclusion and recommended future works are given in chapter 7.

Chapter 2

Literature Review

This chapter expands on the discussion of the current literature addressing the research problem. The following sections seek to review the literature from the perspective of data, architectures, and explainable AI techniques used. It is important to note that a number of studies reviewed in this chapter use a retrospective cohort of data samples collected in the hospital and are not accessible to the research community as open source.

2.1 Data

In a Hammoudi et al.[61] analysis, they wanted to create an efficient classification model that could inevitably determine whether a chest X-ray image was normal, bacterial pneumonia, or viral pneumonia (COVID-19). A total of 5863 children's images from a study by Kermany et al., [27]. Here, the training set, each class contained a uniform distribution of 1345 images, and the test set had 145 images of COVID-19 chest x-rays. These tailored CNN models exploited the CNN backbones of ResNet-34, ResNet-50, DenseNet-169, VGG-19, and dual-use model: Inception ResNetV2-RNN. These models were taught on a training and testing partition, and the DenseNet-169 design performed best with a mean of 95.72 percent. The RNN architecture was particularly sensitive to pneumonia with a 99.3% .

If we observe the literature by Brunese et al.[47], in a three-phase approach, an experimental analysis of 6523 chest X-ray images from different institutions where 250 images were afflicted to patients diagnosed with COVID-19, other lung disorders were represented by 2,753 photos, whereas healthy people were represented by 3,520. They were able to collect chest X-rays of multiple classes by combining the datasets, with Atelectasis, Infiltration, and Pneumonia contributing more than 500 images a piece. The dataset was obtained from institutions from Italy, Korea, Sweden, Australia, and China, thus bringing a non-bias approach to the work. This confirmed the usefulness of the suggested approach, with a detection time of 2.5 seconds and an accuracy of 0.97. VGG-16 was the model that outperformed the others. The Explainable XI method used in this work was Grad-CAM which highlighted the potential pixels that might have contributed to the models' predictive decision.

In a Hussain et al.[48] clinical study, creating an AI imaging analytical model based on portable CXRs to classify COVID-19 lung infection. On 558 chest X-ray images, they extracted texture and morphology features and used five supervised classification technique AI algorithms to differentiate COVID-19 from other conditions. A two-class and multiclass classification was carried out (which comprised COVID-19, normal, viral, and bacterial pneumonia). They used GitHub data from COVID-19 images and non-COVID-19 / standard chest X-rays. There were 250 COVID-19 photographs, 145 non-COVID-19 pictures, 138 bacterial pneumonia images, and 138 routinely diagnosed patients images. Two licensed by the board pulmonary radiologists with substantial expertise assessed these images. They applied a stratified sampling method to the data, with a 70% and 30% ratio training and testing partition. In the statistical analysis, an unpaired two-tailed t-test with unequal variance between groups was used. The ROC curve analysis was used as a performance evaluation metric. The accuracy, sensitivity, and specificity of COVID-19 vs. normal classification for the two-class classification were 100%, 100%, and 100%, respectively. For bacterial pneumonia vs COVID-19 classification, 96.34%, 95.35%, and 97.44% was achieved, respectively. The COVID-19 vs non-COVID-19 viral pneumonia classification achieved a 97.56% 97.44% and 97.67%, respectively. While for the multiclass classification, the accuracy was 79.52% and the AUC achieved was 0.87.

Gu et al. [64], in their work, demonstrated how to use data augmentation and CNNs to improve classification performance. Model overfitting and poor performance are common issues in machine learning and deep learning neural networks, with the goal of maximizing model accuracy while minimizing the loss function.

The CIFAR-10 dataset was used, and the impacts of model overfitting were observed across a variety of architectures. Three datasets where used, MNIST [25], CIFAR-10 [20] and CIFAR-100 [21]. This method showed effectiveness across all deep learning subdivisions including, Natural Language Processing, and Speech Recognition. The CIFAR-10 dataset is a very popular dataset with 50,000 training and 10,000 testing images that are publically available [10]. Their approach was they divided the training into five batches and each batch class contained 5000 images, giving a uniform distribution (no class imbalance) to the dataset. Data augmentation techniques were used as well as hyper-parameter tuning. The 16-layered VGG-16 model trained with a 96% accuracy and tested with a 92% accuracy in both of these techniques. It resulted in losses of less than 0.5 and a classification error rate of less than 8%. Class misclassification rates were lower than 7.5 percent at the time. As overfitting was reduced, this intern assisted in the development of a more reliable neural network model.

Wong et al. [17], their confirmation study used a deep learning system to assess if of computer-aided rating of CXRs of SARS-CoV-2 respiratory disease severity is feasable. They used a COVID-Net-S, which is derived from the COVID-Net network architecture. Here, they learned the network's ability to learn the geographic nature and the opacity extent of COVID-19 on a subset of chest X-rays (CXRs). The dataset was obtained from a patient population from around the world of 267 patients between the ages of 12 - 88 years. They collected 396 chest X-rays of supine, upright, and posterior-anterior positioning. They used Stratified Monte Carlo Cross-Validation experiments for evaluation and yielded R^2 of 0.664 ± 0.044 between predicted geographical extent scores and opacity extent scores with the radiologist scores. The best-performing version of the COVID-Net-S produced R^2 values of 0.739 and 0.741 for predicted and radiologist ratings, respectively. Deep neural networks on the CXR may serve as a useful tool for automated identification and evaluation of SARS-CoV-2 lung illness, according to this study.

So far, we've reviewed literature based on the data we've used. We noted that there were some popular datasets that are readily available to the public and also note that a number of studies reviewed in this chapter use a retrospective cohort data samples collected in the hospital and are not accessible to the research community as open source. We attempted to investigate who used and reviewed the data to ensure its validity, how the data was collected, how the data was distributed, and the classes available in the dataset. In the following section, we will conduct a literature review based on the Deep Learning techniques used.

2.2 Deep Learning Architectures

This section will discuss methods of deep learning that have previously been used to diagnose COVID-19 and other lung diseases.

In a recent study by Sharma et al. [13], COVID-19, non-COVID-19, pneumonia, tuberculosis (TB), and normal chest X-ray shots from a variety of sources were used in the training and development of a robust AI-based classification model, and a TL methodology was used to train and validate 29 various types of AI-based models. The 49 epoch-based model achieved accuracies of 86.26% for the normal class, 65.38% for the COVID-19 class, 42.31% for non-COVID-19, 96.15% for pneumonia, and 73.08% for tuberculosis. The 101 epoch-based model achieved accuracies of 85.71% for the normal class, 70.77% for the COVID-19 class, 51.92% for non-COVID-19, 93.99 for pneumonia, and 74.62% for tuberculosis [13].Apostolopoulos and Mpesiana [40], researchers employed chest X-ray photos from individuals suffering from pneumonia caused by bacteria, proven Covid-19 disease, and normal showing up to diagnose the coronavirus disease autonomously. Transfer Learning was used on 1427 X-ray images from public medical repositories. VGG-19, MobileNet-v2, Inception, Xception, and Inception ResNet-v2 were the CNNs used in this experiment. VGG-19 and MobileNet-v2 performed better with an accuracy of 98.75% for class 2 and 93.48% for class 3 for VGG-19 and an accuracy of 97.40% and 92.85%. MobileNet seemed to outperform VGG-19 when looking at sensitivity and specificity.

In Yoo et al.[65] investigation, they sought to explore if a deep learning-based decision-tree learner could identify COVID-19 from CXR scans. They ran several binary decision trees, the initial of which categorized photos as normal vs abnormal, the following one as normal vs abnormal with tuberculosis symptoms, and the third as normal vs abnormal with COVID-19. The first tree yielded an accuracy of 98%, while the second tree had an 80% accuracy score. The performance in terms of accuracy for the third decision tree was 95%. Likewise in Sethy et al. [59] work. They used a deep learning-based technique to isolate an instance of coronavirus-infected patients. The method had deep feature extraction from the fully connected layer to pass on to the classifier in the training stage. The deep features recieved from each CNN are utilised by SVM classifier and the classification occurs. Eight classifiers based on SVM using deep features of a plathora CNN models were used and these include AlexNet, DenseNet-201, GoogleNet, Inception-v3, ResNet-18, ResNet-50, ResNet-101, VGG-16, VGG-19, XceptionNet, Inceptionresnet-v2. Statistical analysis utilized accuracy, sensitivity, and specificity, and ResNet-50 plus SVM was superior when differentiated to the other classification models with accuracy and specificity of 0.95 and 0.93, respectively. VGG-16 obtained .97 sensitivity.

In Ueda's experiment [55], They present a linear combination method that linearly combines neural networks' statistical pattern recognition-based classifiers. Their approach involved picking a number of neural networks that minimized classification mistakes the most effectively for each class. An ideal classifier would then be created by linearly combining the chosen models. To average the linear weights, they used the minimum error criterion. The combined classifier outperformed the best single classifier in terms of minimizing total classification errors using artificial and actual datasets.

This section reviewed literature based on the Deep Learning techniques that have been used. In the next section, we will review the literature where CNN models are being compared to see the best-performing Models.

In an Asnaoui et al. [16] comparative study, as a quick way to combat the spread of the novel CoronaVirus, an automatic detection and classification system will be implemented. Models such as VGG-16, VGG-19, DenseNet-201, ResNet-50, and Inception_V3 were trained on a dataset of 6087 images, and it was discovered that Inception ResNet_V2 and DenseNet-201 elicited an accuracy of 92.18% and 88.09%, respectively. A paper by Narin et al. [11],as a quick alternative COVID-19 diagnosis option, an automatic detection system was implemented. On a four-class dataset, they proposed five pre-trained CNN classifiers (ResNet-50, ResNet-101, ResNet-152, Inception v3, and Inception-ResNet_v2) and implemented three different binary classifications. A 5-fold cross-validation was utilised and it was seen that pre-trained ResNet-50 proved to be the highest performing model with an accuracy of 99.5% for dataset-1, 96.1% for dataset-2, and 99.7% for dataset-3.

In a Panwar et al. [38] experiment, they proposed nCOVnet, a deep learning neural network-based method for detecting COVID-19, as an alternative and fast screening method. The dataset comprised a total of 337 images. The model predicted COVID-19 correctly with a 97.97% confidence, while it got a 98.69% confidence when classifying other abnormal lung diseases. Khan et al. [**ref26**] in their work, proposed CoroNet which is a deep CNN model based on the Xception model to automatically detect COVID-19. For the 4-class classification, the model achieved an overall accuracy of 89.6% as well as a precision and recall of 93% and 98.2%, respectively. While on the 3-class classification, it produced an accuracy of 95%. Jain et al. [65], in a clinical study, on 6432 Posterior-anterior X-ray scans of the chest obtained from a Kaggle source, three DL-based CNN models were compared. The Xception model outperformed the Inception-v3, Xception, and ResNet-50 models in detecting chest X-ray images, with an accuracy of 97.97%.

2.2.1 Explainable-AI Methods

The study and community of Explainable AI, in recent years, has grown and is very active in developing techniques that are able to mask this "black box" approach. It is defined as having results that are interpretable when object classification occurs especially on the basis of features that humans can understand. Due to the lack of publications regarding misclassification in COVID-19-related chest X-ray classification problems, we have thought of looking at other computer vision problems that use CNNs and seeing how related DL problems were handled. The following paragraphs will speak on the work that has been done in recent years around explaining and being to interpret the models predictions and how they reached certain conclusions.

In Blanco's work [18], they wanted to emphasize the trustworthiness of deep learning models. They identified instances where models would behave in an unexpected fashion, implement an entropy-based method that would assess their reliability, and go on to construct a deep learning-based system that would be able to detect wrong predictions made by itself. They achieved improved accuracy when this method was used through an ensemble of learners.

Kene et al. [74] in geospatial image analysis, used a plethora of algorithms used for the classification of data. Some algorithms were rule-based while some were learning-based. Here, some pixels were misclassified or unclassified even though they got a good classification. They identify that the main reason for misclassifications points to the presents of mixed pixels. They used the subpixel method which gives a better perspective of those pixels and it also varies depending on the type of pixel. They used path radiance, the removal of haze and cloud, and sub-pixel mapping. They then found that feedforward neural networks are not able to reach local minima whereas back propagation neural networks have some advantage but need a lot of time to converge to a minimum value.

Wang, Lina, and Wong [50] in their work presented COVID-Net for the first time, a deep CNN architecture created specifically for identifying COVID-19 cases in chest x-ray images. They investigated how COVID-Net renders forecasts using the GSInquire explainability method in order to acquire an improved comprehension of the critical variables involved in COVID-19 cases, as well as to evaluate COVID-Net in an accountable and honest manner to support its claim that it renders selections based on appropriate data from X-rays.

Pavlova et al. [51] introduced COVID-Net CXR-2. When compared to the previous COVID-Net, COVID-Net CXR-2 is an improved deep CNN design for COVID-19 identification of chest X-Ray pictures employing a bigger number and more varied patient population. The network attained a 95.5 percent sensitivity and a 97.0 percent positive predictive value. This was transparently and responsibly verified, and to learn more about the models' decision-making processes, a performance validation based on explainability was applied.

In a Hendricks and Gimpel experiment [26], they considered looking at two related problems of detecting if an example is misclassified. They looked into if a baseline that utilizes probabilities from the softmax distribution would be able to identify correctly classified examples and those that are erroneously classified and out of distribution. The correctly classified examples show greater maximum softmax probabilities. Three datasets where used, MNIST [25], CIFAR-10 [20] and CIFAR-100 [21]. This method showed effectiveness across all deep learning subdivisions including, natural language processing and speech recognition.

Through our readings and observations of previous work and literature, we can see and adopt a number of methods and techniques for our study. We were able to identify the COVID-19 data in Cohen et al.'s study has been widely used and is unbiased with these relevant and highest number of COVID-19 samples, so we are going to incorporate that dataset for training our baseline model based on the literature that uses shallow and deep learning mechanisms. We also identified that for shallow networks, VGG-19 has shown incredible performance in other problems and will be adopted in this research report. A plethora of deep networks such as InceptionNet also were good classifiers in other literature. But ResNet-50 was adopted in this research for simplicity and no duplication. We also looked into a network that was pre-trained on COVID-19 images where InceptionNet is the base model used. Transfer learning is then applied to use the weights on the state-of-the-art neural network, CoroNet. 5-fold cross-validation is adopted in the study such that the model is trained on every instance available on the dataset. The next Chapter seeks to explain the CNN architecture in more detail as well as the models that have been adopted in this research report. We will also explain the evaluation metrics used in classification problems.

Chapter 3

Methods

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a class found under the umbrella term deep neural networks together with commonly known Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) to mention a few. CNN is more commonly used in image recognition problems [11]. When an image is given as an input, the image is converted into a format that the computer can recognize and process, a $n \times m \times l$ matrix. By processing the matrix, the system decides which shots go with which labels. During the training, phase is where it learns and labels them into their appropriate classes. When a new (unseen) image is fed into CNN, the goal is for the image to be classified or predicted into its appropriate class. To perform operations effectively, a CNN is composed of three layers. These layers are the convolutional layer, the pooling layer, and the fully connected layer, which will be discussed further below.

3.1.1 Convolutional Layer

The base layer of a CNN, as its known, enables the features to determine the pattern the image follows or has. A filter matrix also termed the kernel, is convoluted on the image using the convolution operation. Within this stage, an element-wise multiplication occurs between the 3×3 or any arbitrary-sized matrix kernel and the image to extract features in the pattern that are classified as low and high. A parameter called stride indicates the steps for shifting over the input matrix. This layer is denoted as:

$$x_j^l = f(\sum_{a=1}^{N} w_j^{l-1} \times y_a^{l-1} + b_j^l) \tag{3.1}$$

where x_j^l is a feature map of number j-th in layer l, w_j^{l-1} is the $j-th$ kernels in the layer before l, this is represented with l-1, y_a^{l-1} is the feature map of the a-the value in the $l-1-th$ layer. The bias is denoted as b_j^l and is of the $j-th$ map of characteristics in the lth level, N is the number of attributes in the layer that preceding l, denoted as l-1. A vector convolution process takes place in this layer.

3.1.2 Pooling Layer

This layer targets the feature maps created and is applied to reduce the feature map's quantity and parameters of the network. This is done by implementing mathematical computations. It reduces the size of the representation to speed up computation and also to make its features in the dataset more robust. This layer applies different types of poolings such as max-pooling and global average pooling. The maximum value is selected when using the max-pooling, and this is done on each feature map to reduce the output layers, while the global average pooling reduces the dimensionality of the data to a single dimension. A dropout layer is the intermediate layer used which is a preventative layer of the network overfitting and diverging.

3.1.3 Fully Connected Layer

This layer is notably the most important layer. It is also the last layer for CNN operational work. It works in the same way as multi-layer perception and employs the Rectified Linear Unit (ReLU) activation function. In the final layer of the fully connected layer, the softmax activation function predicts the output picture probability. Mathematically, these activation functions are denoted as follows:

$$ReLU(x) = \{^{0, x \leq 0}_{x, x \geq 0} \tag{3.2}$$

$$Softmax(x_i) = \frac{e^{x_j}}{\sum_{y=1}^{m} e^{x_y}} \tag{3.3}$$

3.2 Models Selected

In this study, a TL experiment was conducted using relatively new tailored model architectures. TL's benefit is that it enables data training with fewer datasets and with lower computational costs. [71]. The aim of this was to see if these new models outperform antecedent models. The state-of-the-art model is CoroNet [**ref26**][66].

3.2.1 Coro-Net

CoroNet is a CNN architecture created especially for the detection of COVID-19 in chest X-ray images. It is based on the extreme variation of the Inception model called the Xception CNN architecture [**ref26**]. The ImageNet dataset is used to pre-train it [50][**ref26**] and after that, it's trained on the COVIDx dataset. Xception employs depthwise separable convolution layers in place of conventional convolutions. These swap out the $n \times n \times k$ convolution process for $1 \times 1 \times k$ point-wise convolution operation, which is followed by $n \times n$ channel-wise convolution operation. Residual connections, commonly referred to as "skip connections," are the primary means by which gradients traverse a network. Without going through the non-linear activation function, this is accomplished. The original weight layer output series can be combined with the residual connections and then sent via a non-linear activation function. A dropout layer is added after the base Xception model and two fully connected layers are added at the end.

In this study, we will also integrate a method for trained network deep feature extraction. While some designs are easier than others, other complex designs demand GPU. During preprocessing, the input size of X-ray pictures will be altered to fit the targeted pre-trained networks and match the inputs, The baseline architecture of choice for this project is ResNet-50 and VGG-19.

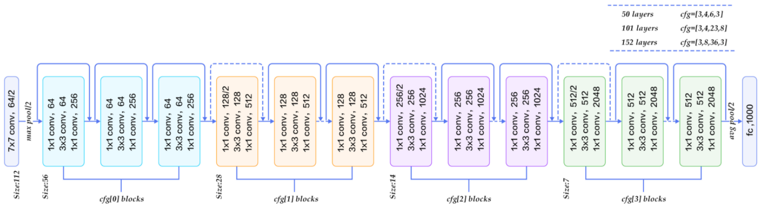

FIGURE 3.1: A diagram showing the ResNet-50 architecture.

3.2.2 ResNet-50

ResNet-50 is known for adding shortcuts between layers to solve issues The Residual Neural Network algorithm is an enhanced version of the CNN model. This quality prevents distortion that may occur as the network becomes complicated and more in-depth. As its name suggests, this model has 50 layers that are trained on the Imagenet dataset [11]. It leverages residual design principles and a lightweight design pattern with moderate architectural diversity but doesn't leverage a lightweight PEPX design pattern, to its comparison COVID-Net [50]. In recent years, it has produced some of the most innovative work in the field of deep learning and computer vision. ResNet allows for training with hundreds or even thousands of layers and still keeping high performance. Highway Network developed gated shortcut connections, and it uses shortcut connections. The amount of information that can pass through the shortcut is controlled by these parameterized gates.A parameterized forget gate in the Long Term Short Memory (LSTM) cell utilizes a similar principle to control how much data goes to the next time step. ResNet can thus be thought of as a subset of Highway Network.

3.2.3 VGG-19

VGG-19 is a shallow deep neural network and a version of the VGG model, initially created as a group name Visual Geometry Group. It consists of 19 layers (16 convolution layers, three fully connected layers, five maxpool layers, and one softmax Layer) and consists of 19.6 billion flops. Different variants of VGG are available, Namely VGG-11 and VGG-16. VGG is a successor of AlexNet.

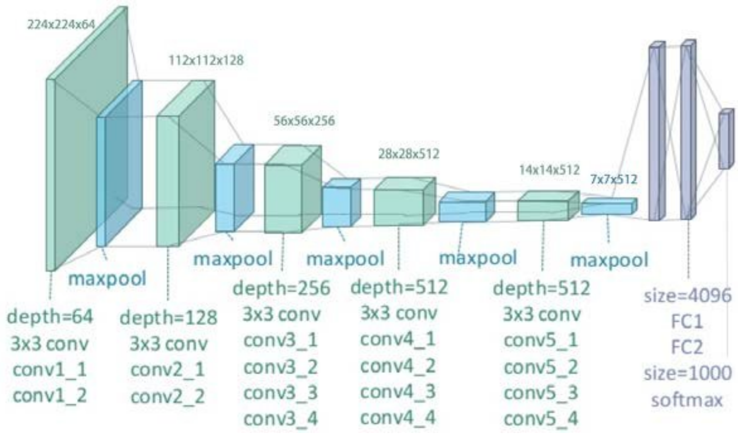

FIGURE 3.2: A diagram showing the VGG-19 architecture.

3.3 Analysis

The proposed model will be evaluated using the following evaluation metrics. The Deep Feature Extraction approach included accuracy, precision, recall, specificity, and F-score. While the evaluation of the Transfer Learning approach included the following: Accuracy, Precision, Sensitivity, and Specificity. These measures are given by the following equations, where FN stands for False Negative, TP stands for True Positive, TN stands for True Negative, and FP is for False Positive.

$$Accuracy = \frac{correctly\ classified\ images}{totalnumber\ of\ images} = \frac{TN + TP}{FP + FN + TP + TN} \tag{3.4}$$

$$Precision = \frac{Sum\ of\ all\ TP}{Sum\ of\ all\ TP + FP} = \frac{TP}{TP + FP} \tag{3.5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.7}$$

$$Recall = \frac{Sum\ of\ all\ TP}{Sum\ of\ all\ TP + FN} \tag{3.8}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.9}$$

3.4 Explainable AI Techniques

This section seeks to give an overview and explanation of how these XAI methods work for the readers that may be interested to know more about these. Readers that are familiar with this may skip the following section and proceed to the Methodology used in this study.

3.4.1 Integrated Gradients

This approach is based on two fundamental axioms: implementation invariance and sensitivity. Sensitivity relates to the number of greater than zero attributions attributed to each intake and baseline that vary in a single attribute but create distinct forecasts. To calculate, we'd need to start with a baseline image and then build a sequence of images that interpolated from the baseline image to the actual image. When two functionally equivalent networks have identical attributions for the same input image and baseline image, implementation invariance is satisfied [6]. This method combines gradients with each other $\frac{\partial f(x)}{\partial x_i}$ for model f, calculated down the path from the instance of interest x to the highlighted observation, which is the baseline x^*. Integrated gradients are defined as follows:

$$IntegratedGrads_i(x) = (x_i - x_i^*) \int_{\alpha=0}^{1} \frac{\partial f(x^* + \alpha(x - x^*))}{\partial x_i} d\alpha \tag{3.10}$$

Four gradient-based methods have been studied in the past decade. These include Saliency, Integrated Gradience, GradientShap, and FusionGrad [4]. Saliency alludes to features that are unique to images such as pixels or the resolution of images. These features try to delineate the visually prominent and accentuate locations in an image and saliency maps the topographical representation [60]. We will make use of Integrated Gradients, a technique that aggregates gradients along a linear path between the sample and the baseline to compute feature attribution. The two chosen theoretical deliverable promises that it favors for the application of this strategy are presented. To begin with, it is delicate. This approach assigns a non-zero value to each intake and baseline that vary in one way but give distinct forecasts. Second, it is implementation invariant, which implies that if both models are used similarly (same result given the same input), their characteristic attribution will be identical regardless of network design. [8].

Chapter 4

Research Methodology

4.1 Research design

This section intends to discuss the implementation of the research design. It consists of five subsections which will give a deep description of the tools and techniques that were used in this research. This section will start off by discussing how the data was collected and pre-processed, a discussion on the models that were used, and matrices that were applied to evaluate the model's performance are explained, we will also discuss Explainable-AI techniques that explain how misclassification might have occurred. Finally, a description of the tools used to carry out this experiment is explained.

4.1.1 Model Implementation Environment

All experiments were performed in a Jupyter Notebook using the Python (version 3.9) Programming Language with a Windows 11 CORE i7, 16 GB RAM, and LENOVO ThinkPad. Python programming language is an advantage to training the proposed transfer learning models as well as the deep feature extraction processes as the Tensorflow framework and Keras libraries offer the tools needed. Since DL requires high-performance computing, cloud services found online with Graphic Processing Unit(GPU), Central Processing Unit(CPU), and Tensor Processing Unit(TPU) were deployed for a sound execution of the experiments.

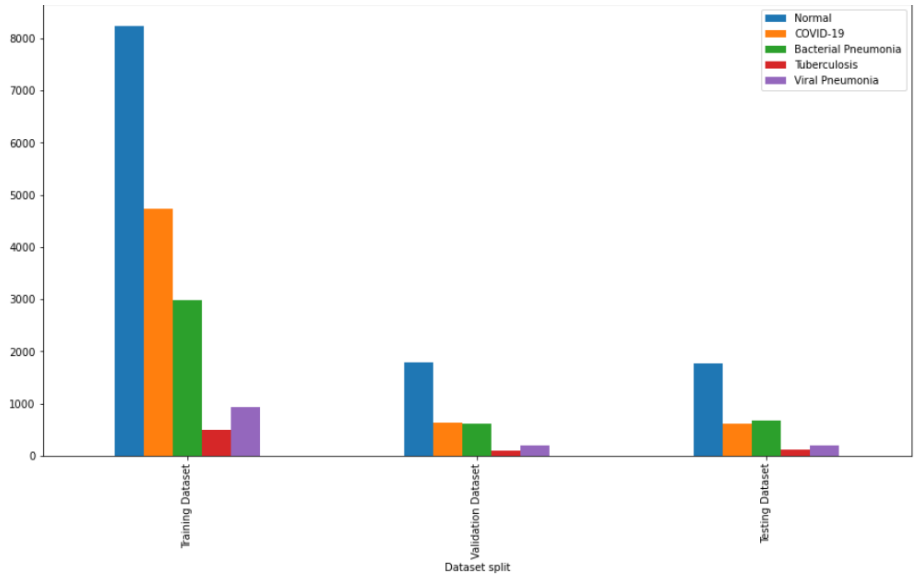

FIGURE 4.1: Sample distribution of the dataset. This is to show the distribution of the split into independent datasets with a 70%, 15% and 15% split for the training set, validation set, and testing set respectively.

4.2 Dataset

The 5-class COVID-19 dataset was constructed and taken from Four different sources. This dataset contains Posteroanterior(PA) chest X-ray images and any other view and CT images were excluded from the sample. These images are set to resemble real-world conditions to try to capture the challenging qualities found in real-world qualities. These included low-quality images and differently aligned images to represent the diversity seen in real-world cases. The COVID-19 cases were acquired from Cohen et al. dataset [22], an open-source GitHub Repository. This dataset is said to have the largest count of X-ray images related to COVID-19 [3]. Chest X-ray images related to bacterial and viral Pneumonia were obtained from Kaggle, another popular Online repository [23]. The Tuberculosis cases were selected from the Kaggle TB dataset [24]. The images from typical (healthy) X-rays of the chest were chosen from the "ChestX-ray8" database [42]. All the images were resized to 225 x 225 pixel size in the dataset.

From the total number of images, the IDL groups were classified as a 5-class dataset with the following classes: Normal, COVID-19, Bacterial-Pneumonia, Viral-Pneumonia, and Tuberculosis. The dataset was then randomly split into independent datasets with a 70%, 15%, and 15% split for the training set, validation set, and testing set respectively.

4.2.1 Data Preproccessing

After the data is obtained, any notable bias in the samples was dealt with by the use of data augmentation as stewed data for the model will show an unbiased performance towards a certain class. These techniques are used to regularize neural network models that work with images. The synthetic images were created by applying various transformations to the original images from the labeled training dataset. These augmentation techniques include (1) flipping images vertically and horizontally; (2) rotating some acute degrees; (3) rescaling inward and outward; (4) cropping images; (5) translation by width and height shifts; (6) whitening; (7) shearing; and (9) zooming and adding to prevent overfitting, adding 10% Gaussian noise.

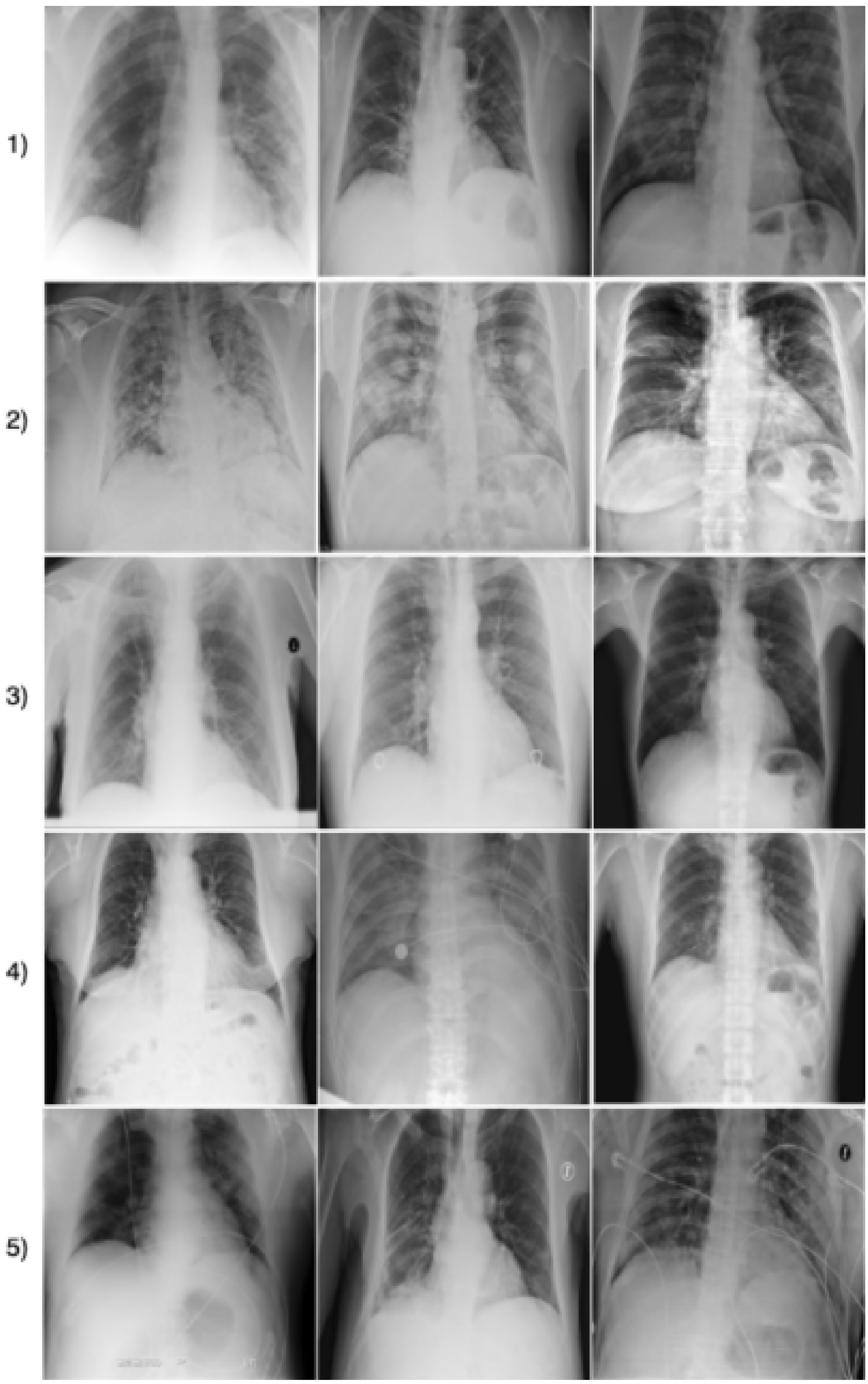

FIGURE 4.2: Sample Images from the 5-class COVID-19 Dataset, which comprises 24 898 chest X-ray images. Class labels: 1) COVID-19 chest X-rays, 2) Bacterial-Pneumonia chest X-rays, 3) Normal chest X-rays, 4) Viral-Pneumonia chest X-rays, 5) Tuberculosis infected chest X-rays

4.2.2 Experimental setup

Base Model

The current work investigated the difference in performance between "state-of-the-art" models utilizing transfer learning and those that use deep feature extraction in order to form a base model that was used in the second and third sections of the experiments. Three models were implemented in total. ResNet-50, VGG-19, and CoroNet will be the models used in this study. The first two models were built from the ground up, whereas the latter had weights pre-trained on the ImageNet dataset and then trained on the dataset created for this study. The following hyper-parameters that were utilized for the training process are as follows: Learning rate = 1e-4, Epochs = 25, with the Batch size = 128 and the Factor used = 0.7. All three models were built and evaluated using the Keras DL library with Tensor-Flow back-end. The best-performing model trained was carried through to the next section of the experiment.

K-fold Cross Validation

In the second part of the experiment, the most effective model was then retrained on the original samples and the augmented samples. A k value of 5 was given to a k-fold cross-validation method that was implemented. The technique known as "K-fold cross-validation" merely repeats the cross-validation process a number of times and then returns the average results across all of the specified K folds. It seeks to offer a means of enhancing a model's predicted performance. The Data was divided into training, validation, and testing sets, with the testing set being removed from the process in order to effectively inject variance into the model. 20 percent of the training set was divided into validation sets, and 20 percent of each fold was randomly chosen to act as the validation set.

Model Evaluation

The model's performance will be assessed using a confusion matrix, with accuracy, the F1-score, and sensitivity being useful for the training, validation, and testing sets. Individual class predictions are compared using the characteristics derived from the confusion matrices.

Explainable AI method: Integrated Gradients

The third section of the experiment is due to wanting to understand how these models made their prediction. In this part of the experiment, the Explainable AI technique of choice will be Integrated Gradients (IG) where we will look into its attributes that will help us explain misclassification by these models. Python tools focused on deep learning for images, and the adoption of the concept of saliency in images was observed.

4.3 Limitations

Previous literature's main challenges are related to having a limited count of images to make a dataset that CNN requires. This tends to lead to an unbalanced dataset and we plan to look at augmentation techniques as well as explore other solutions. Another issue is the limited computational resources available, such as GPUs, to train newly proposed CNN models and those that are pre-trained already. A limitation also identified is that the separability in a class of COVID-19 Deep Features is not known especially where various diseases are in consideration.

In this section, we described the methodology that was employed in this study. We mentioned the research design and what equipment and tools were used throughout the study, and we then went on to describe the data, its distribution, and preprocessing methods applied to the data. We then took you through the experimental pipeline developed to reach the objectives set out and concluded with the limitations that were faced. The next chapter will lay out the results and discuss the findings of this study.

Chapter 5

Results and Discussion

5.1 Experiment Results

5.1.1 Base Models Performance

In this study, two experiments were performed. The base model experiment and the k-fold cross-validation method. In the first (base model) experiment, we combined different datasets to form a multi-class classification with 5 different uncorrelated classes (Normal, COVID-19, Tuberculosis, Viral, and Bacterial Pneumonia) to obtain a robust outcome. This was performed on two different pre-trained models, ResNet-50 and VGG-19 as well as a state-of-the-art model, CoroNet. In the second experiment (k-fold cross-validation method), we presented the best-performing neural network of the three, in this case: CoroNet and further training were performed to understand the misclassification that occurred by means of explainability and interpretability. For the results of the initial experiment, their training and validation accuracy and loss are plotted to evaluate the efficacy of the models put forth for the experiment. The accuracy of the model presented is described in the following table. Table 5.1 shows the model performances obtained from images trained on the original dataset. This illustrates how well the model

TABLE 5.1: Base Model Performance Comparison on Original dataset

Results obtained from the first experimentation				
Algorithm	ACC(%)	REC(%)	PREC(%)	F1(%)
ResNet-50	85	80	91	85
VGG-19	45	3	64	6
CoroNet	98	98	97	97

performed on the training and testing set.

The models were trained with varied layer lengths and architectural designs, and the accuracy and loss values during the training and validation processes were acquired from the plots below (Figure 5.1 - 5.3). CoroNet was clearly the best-performing model when compared to the others. The CoroNet model achieves a reduced loss value when compared with the loss values of the other models. The multiclass classification results for these models were captured and represented in the table below. These were trained on the Normal dataset, which will be referred to as the original dataset. The different deep-learning pre-trained models such as ResNet50 and VGG19 were used in this experiment because of their performance in the dataset that had Pneumonia as the negative class [50]. The other reason is that these two DNN architectures don't possess the key defining traits that CoroNet contains. CoroNet achieved an accuracy of 98% on the unprocessed data. The high performance may be attributed to the use of transfer learning as well as domain-specific transfer. While ResNet-50 and VGG-19 were both pre-trained on the ImageNet dataset, CoroNet has been pre-trained on COVID-19 images and is thus familiar with the features that COVID-19-related chest X-rays may contain, and thus the model is able to learn. The quantitative performance of these models can be seen in Table 5.1. This then allowed us to continue the experimentation with CoroNet architecture.

The confusion matrix shows the models' performances in correctly classifying or misclassifying these images. A number of interesting observations can be deduced from the confusion matrix of CoroNet. These confusion matrices can be observed in Figures 5.4 to 5.6. The confusion matrix of the CoroNet architecture illustrates a great representation of the classifications being correct and accurate. In Figure 5.6 we see that a large number of the datasets were correctly classified, which can be observed with the diagonal matrix having high numbers compared to other boxes. When compared to Figures 5.4 and 5.5, the diagonal line is not present. The CoroNet architecture only falsely predicted 11 images as original which are actually COVID-19 infected and only one image was falsely predicted as containing TB and not COVID-19. In Figure 5.4, if we observe Actual COVID-19 and Predicted

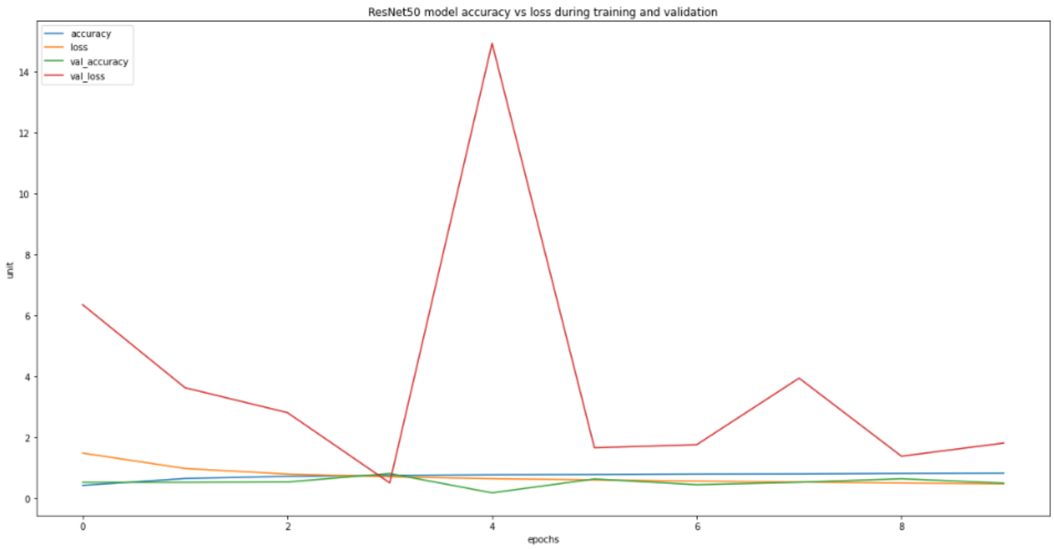

FIGURE 5.1: Accuracy and loss of the training and validation set on ResNet50 model

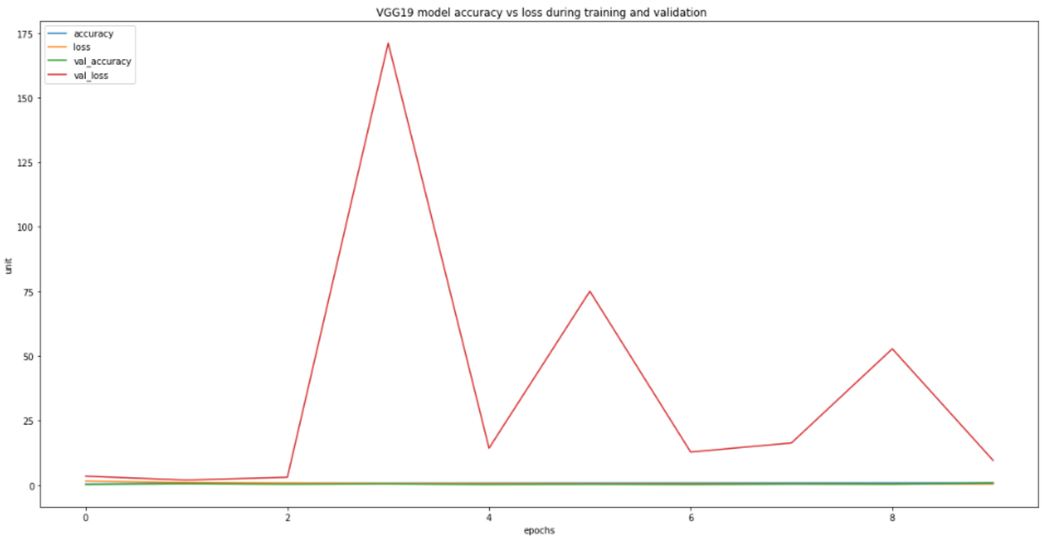

FIGURE 5.2: Accuracy and loss of the training and validation set on VGG19 model

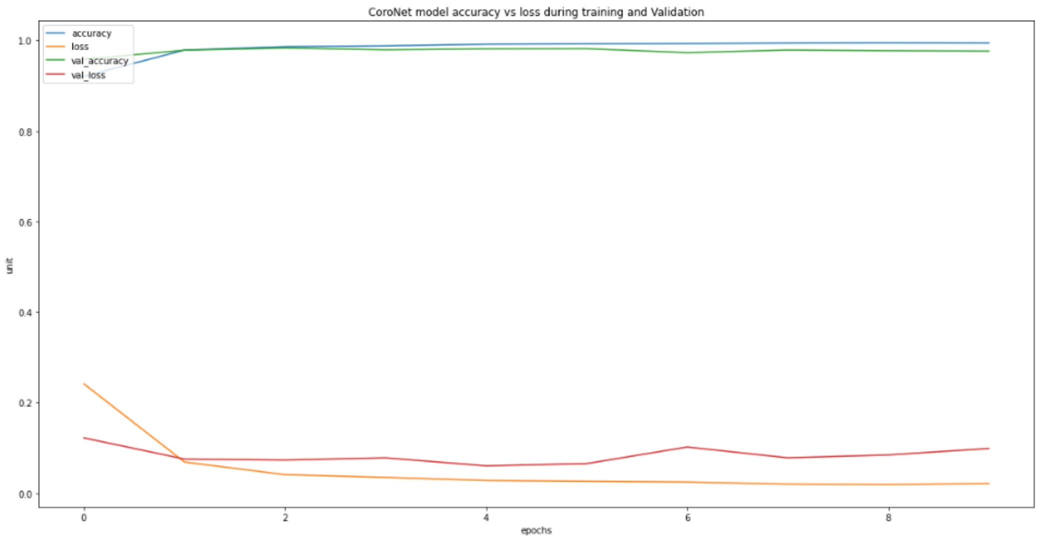

FIGURE 5.3: Accuracy and loss of the training and validation sets on CoroNet model

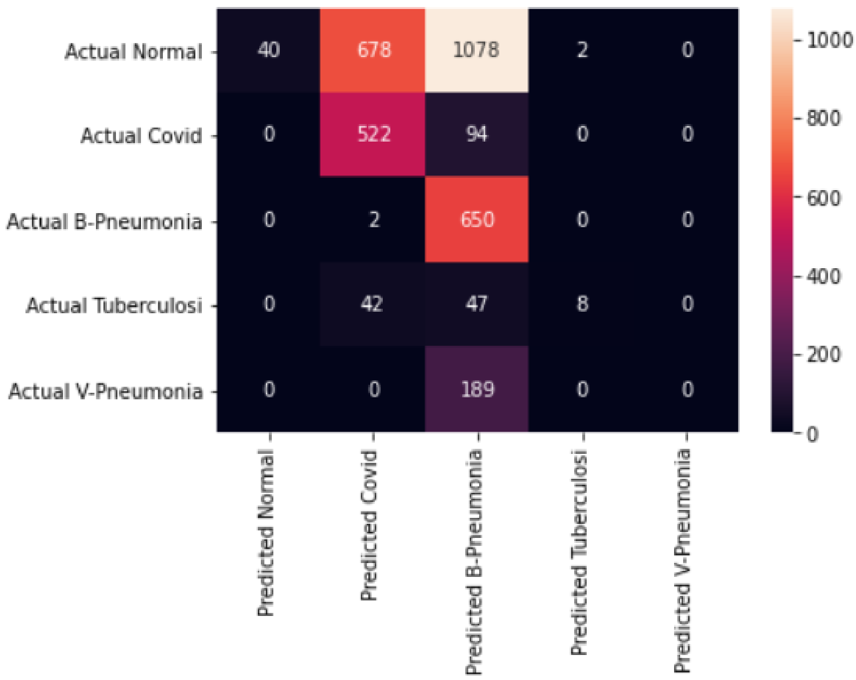

FIGURE 5.4: On the original dataset, a confusion matrix for the ResNet50 architecture.

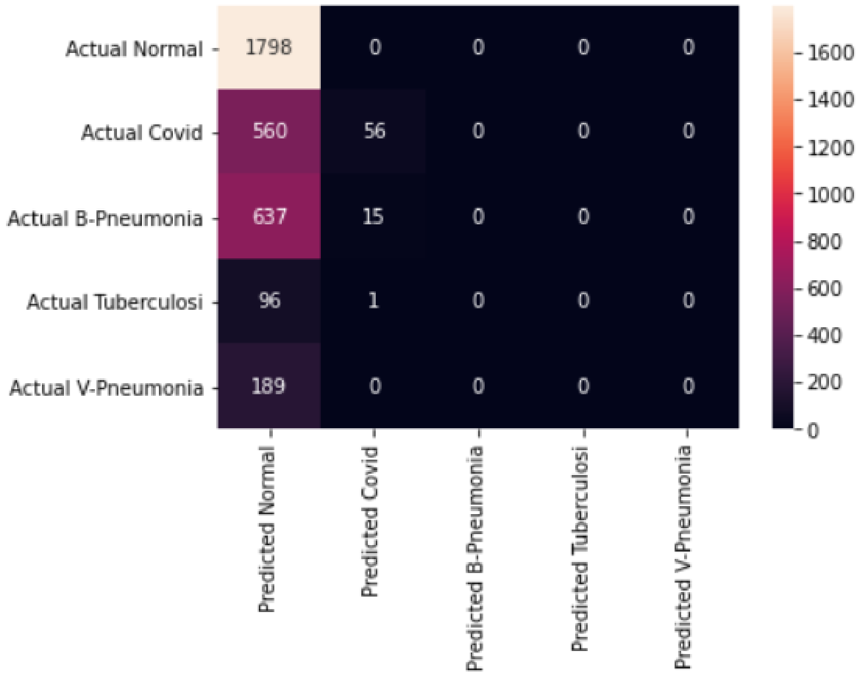

FIGURE 5.5: On the original dataset, a confusion matrix for the VGG19 architecture.

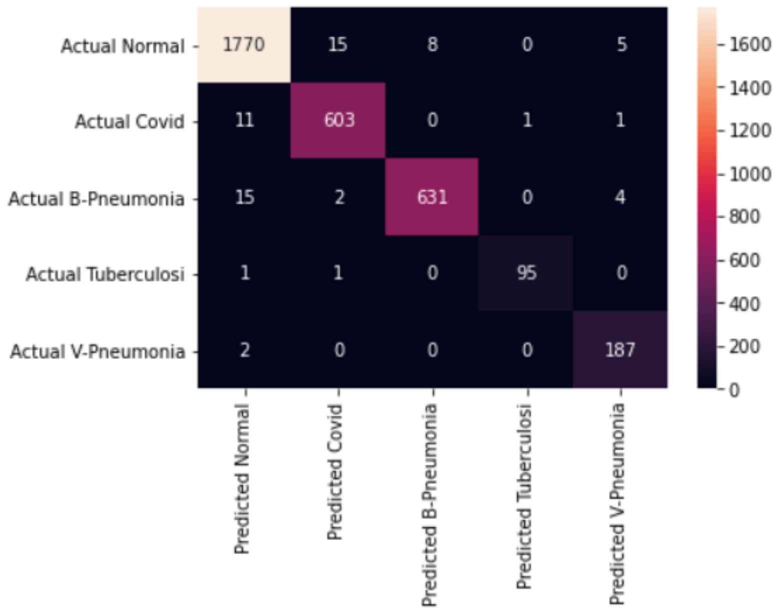

FIGURE 5.6: On the original dataset, a confusion matrix for the CoroNet architecture.

COVID-19, we will Notice some correct classifications but there are some misclassifications for actual COVID-19 predicted as Pneumonia. As we have said, a diagonal matrix is ideal for us in these representations but besides the prediction of COVID-19 and Bacterial Pneumonia, All had significantly failed. Looking into Figure 5.5, only the normal instances were predicted correctly whereas the other classes failed to be correctly classified.

The misclassification can be attributed to a plethora of cases. If we look at The distribution of the dataset. The training, validation, and testing sets were distributed using the 70%, 15% 15% split, respectively. Within the validation set, the Normal class, COVID-19, Bacterial Pneumonia, Tuberculosis, and Viral Pneumonia had 1755, 651, 630, 213, and 102 cases respectively. Due to this class imbalance, both ResNet-50 (Figure 5.4) and VGG-19 (Figure 5.5), were out-perforated and had a very high false positive rate. The class imbance would have favoured the classification of the classes with the highest cases. This would have caused the model to learn and familiarise itself with the Normal case patterns and would've been the reason the other classes were incorrectly predicted. This can be seen in Figure 5.5, where VGG-19 predicted all of the classes as Normal classes. Hence, the poor performance. The other cause might be due to the similarity in which the images look. When comparing the samples that were actually normal but were predicted as bacterial Pneumonia and actual bacterial Pneumonia, it may appear that the images look quite similar, the model mistakenly predicting Normal images as Pneumonia.

5.1.2 K-fold Cross Validation

In the second experiment, CoroNet was then presented as the best-performing neural network of the three, and further trained and tested to understand the misclassification that occurred. The network was trained on a training and validation set of both the original dataset and the augmented dataset and tested on the testing set using 5-fold cross-validation. The results obtained are evaluated and tabulated in Table 5.2 and Table 5.3. It can be said that the accuracy values per fold do well as they are fluctuating in the 80% margin on the original dataset and test accuracy of 84%. Specificity on the other hand received 89%. For the augmented dataset, the

TABLE 5.2: CoroNet 5-fold Cross-Validation performance for the original dataset:

Results for CoroNet architecture on original dataset					
Fold	ACC(%)	SPE(%)	SEN(%)	REC(%)	F1(%)
1	88.09	95.19	71.29	88.03	87.87
2	88.24	94.75	70.33	88.14	88.02
3	84.96	93.31	74.22	85.38	84.30
4	87.72	93.55	81.08	88.11	87.42
5	89.44	91.26	85.90	89.81	89.43
Test	84.02	89.87	70.95	84.12	84.05

TABLE 5.3: CoroNet 5-fold Cross-Validation performance for the augmented dataset:

Results for CoroNet architecture on augmented dataset					
Fold	ACC(%)	SPE(%)	SEN(%)	Prec(%)	F1(%)
1	85.21	96.85	68.64	86.13	83.86
2	88.79	95.70	73.46	88.81	88.53
3	85.61	91.82	78.07	86.15	85.28
4	86.40	94.05	76.51	86.85	85.74
5	87.42	92.02	83.37	87.94	87.29
Test	82.00	89.18	71.13	82.49	81.88

test accuracy was 82%, which is lower compared to the performance observed from the original dataset.

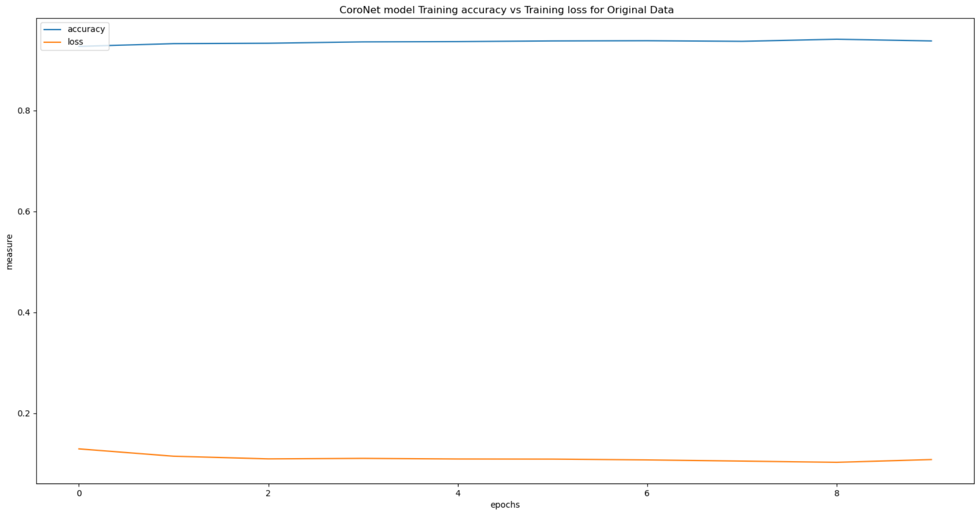

FIGURE 5.7: Training Accuracy and Loss of the CoroNet model for Fold-4 of the experiment on the original dataset

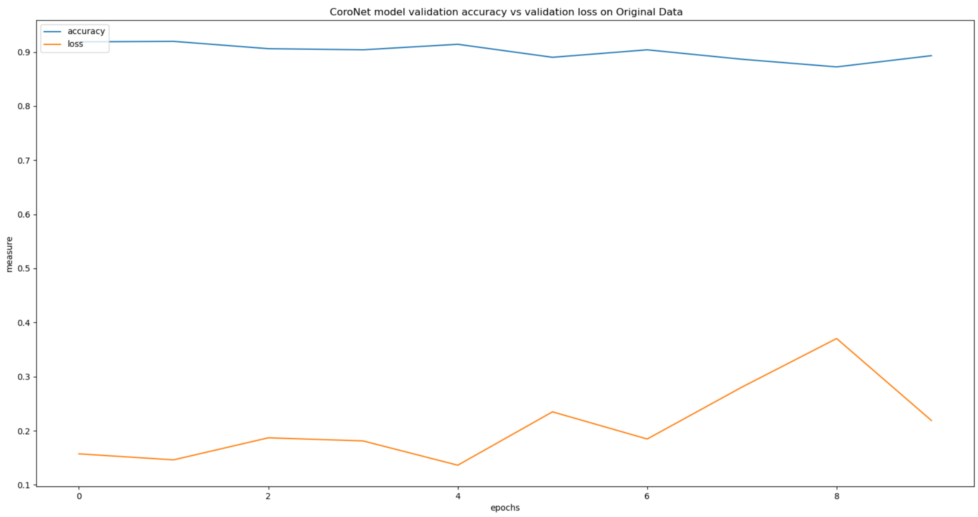

FIGURE 5.8: Validation Accuracy and Loss of the CoroNet model for Fold-4 of the experiment on the original dataset

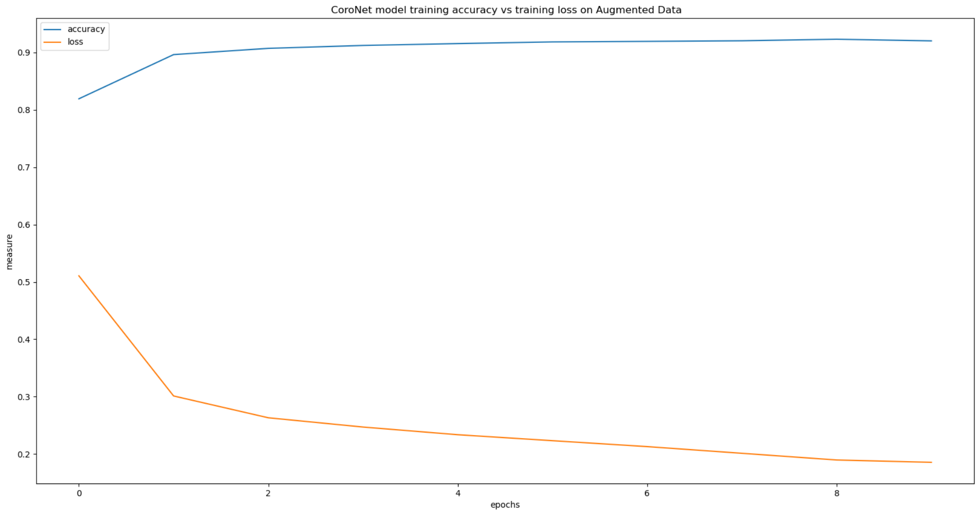

FIGURE 5.9: Training Accuracy and Loss of the CoroNet model for Fold-4 of the experiment on the augmented dataset

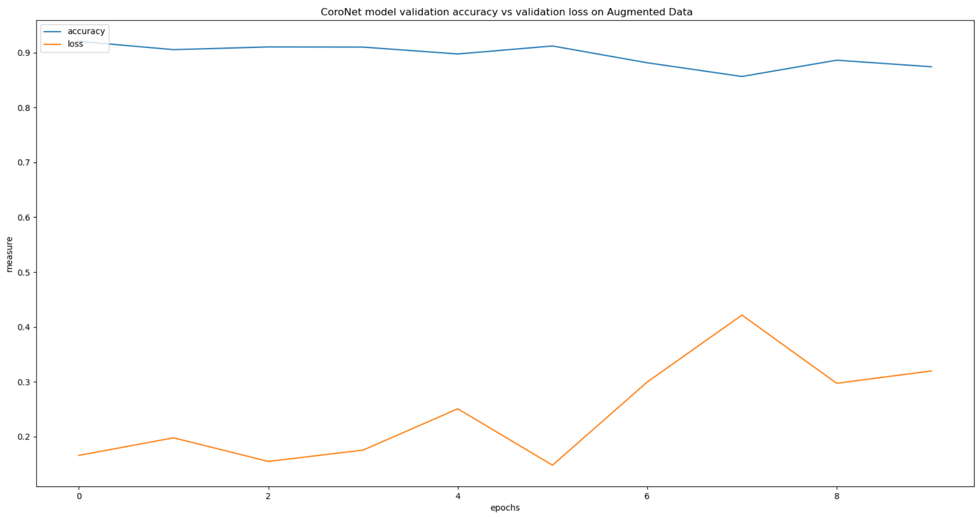

FIGURE 5.10: Validation Accuracy and Loss of the CoroNet model for Fold-4 of the experiment on the augmented dataset

5.1.3 Explainable-AI using integrated gradients

The results from the conducted explainability-driven performance validation (see Figure 5.11 and Figure 5.12) show clinically relevant visuals of the lungs indicating why there was a misclassification of these images and how it was leveraged in the decision-making process. As mentioned, Integrated Gradients was used as the Explainable-AI method, which highlights the pixels that the model used to learn and make predictions from [36]. Although Integrated Gradient has not been explicitly been implemented in a case of predicting COVID-19 from other interstitial lung diseases, It was applied and Compared to Guideded Integrated gradience(IG) on retina image to assist in diagnosing diabetic retinopathy, They found that Guided IG clustered it attribution around the predicted class object with less noise around other parts of the image [9]. However, GRAD-CAM, which was applied to chest X-ray classification, was used to visually validate where the network was observing the correct patterns in Chest X-ray images and activations around the image. GRAD-CAM works by looking at the final convolutional layer in the network and examining the gradient information flowing to that layer. As a result, the activation maps generated on the overlap on the chest X-ray were thought to be useful to radiologists as they were able to localize the areas on the X-ray to investigate. The areas mapped showed a probability of the predicted label to be higher [47].

If we observe our work, in Figure 5.11, the model accidentally classified a COVID-19-labeled image and classified it as a normal labeled image due to the pixels being highlighted by the Model. If observed closely, a triangle shape can be visually seen from the masked normal image and the masked COVID-19 image. This can be how the model was able to make this mistake because structurally these images are the same and highlight more or less the same region on the image. Figure 5.12, also has a similar trait as the model also highlights the pixels that run in the center down to the bottom right of the image. Other work closely related to IG was proposed in [60], where attributions are used in order to help identify three question-answer models better than CNN models.

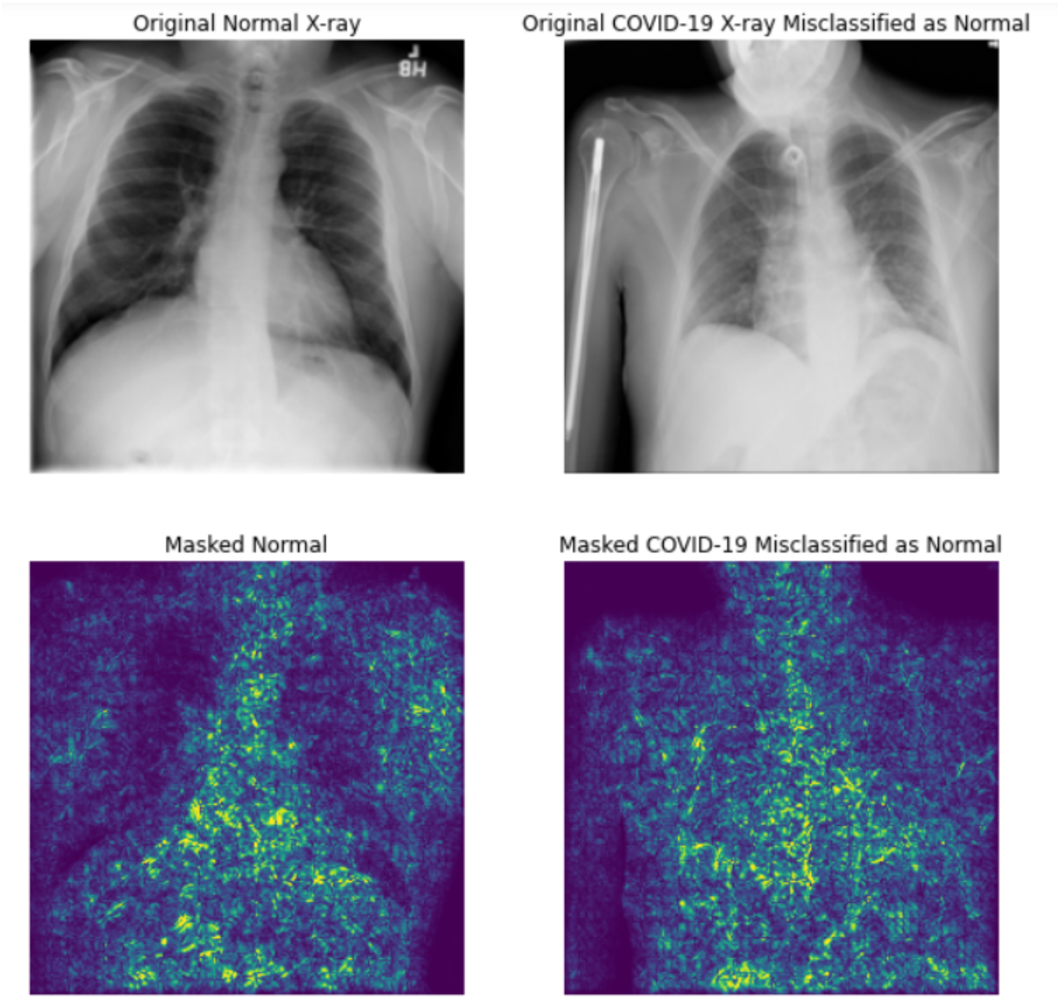

FIGURE 5.11: Integrated gradients showing misclassified COVID-19 predicted as Normal or non-infected patients and why there was a misclassification.

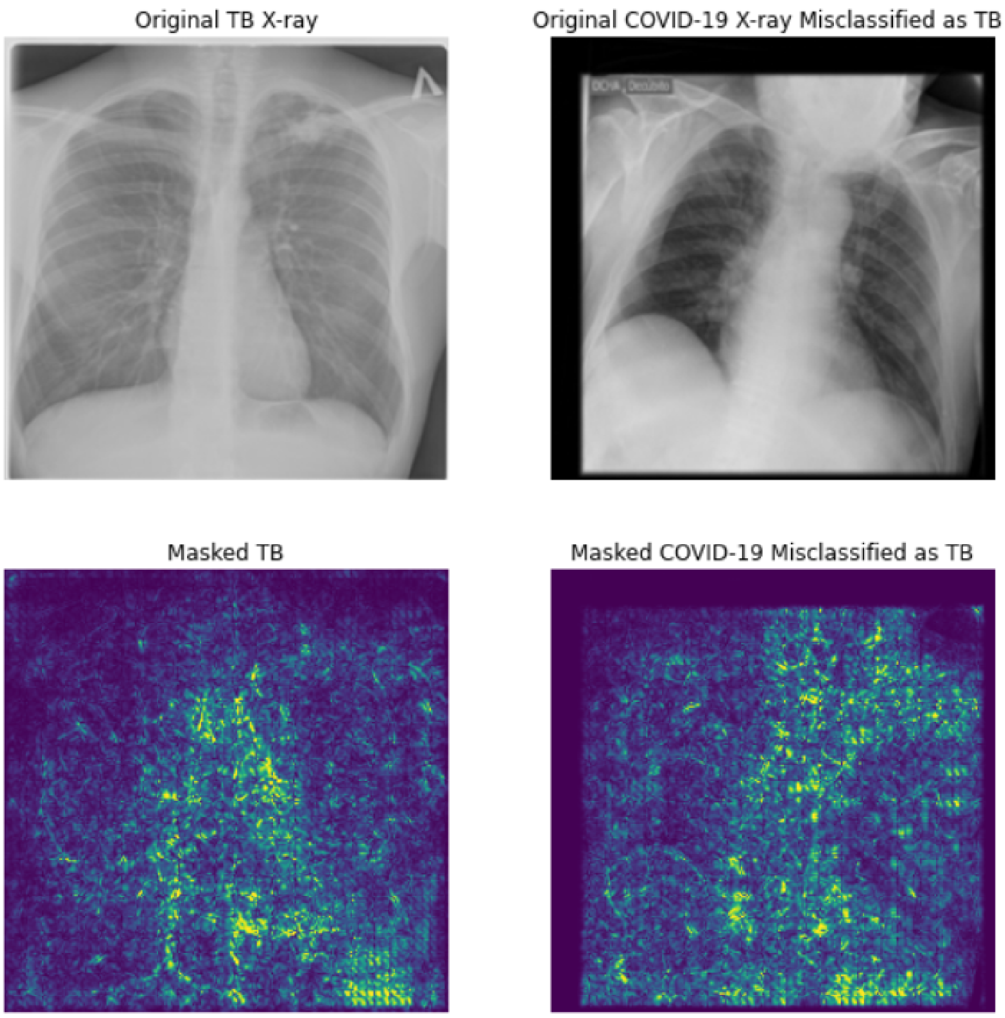

FIGURE 5.12: Integrated gradients showing misclassified COVID-19 predicted as Tuberculosis and why the was a misclassification.

5.2 Discussion

Medical Imaging has played a crucial role in combatting not only COVID-19 infection but in a plethora of other disciplines in medicine alone [12], [73], [2]. Deep learning has been used in several examples both in the medical space as well as other fields [63], [70], [53], [34]. The focus of this research report has been on medical imaging, specifically focusing on Interstitial Lung Disorders. We identified and observed infections that are common as well as mentioned in past research. Due to recent events, COVID-19 has taken the world by storm, killing millions of people all over the world. In this study, we implement the Inception architecture-based deep model to detect COVID-19 among other lung diseases. To understand why these models misclassify images, it is important to distinguish between images of viral pneumonia-infected X-rays and bacterial Pneumonia infected X-rays, COVID-19 infected and Non-COVID-19 infected X-rays as well as Tuberculosis and COVID-19 infected X-rays.

In a study by Khan et al. [7], they proposed the CoroNet model on a 4-class dataset of 1251 images, which was taken from two sources. The data was partitioned using an 80% spit for training and the remaining 20% was for the testing set. They obtained an accuracy score of 89.5%, sensitivity of 100%, precision of 97%, and f1-score of 98%. This study, however, did not include Tuberculosis CXRs. Another Study that also used CoroNet was khobahi et al. which acquired data from four different sources. Here, they compared COVID-19, pneumonia, and normal CXR images by using a 90%:10% dataset partition to achieve a 93.5 % accuracy, 93.5 % for sensitivity, 93.63% precision, and f1-score of 93.51%. Here, they worked with a very larger sample size than most (18,529 samples) but the number of COVID-19 cases amounted to only 99. Their CoroNet model also used Auto-encoders as part of its architecture and training technique [67].

A number of 3-class classification problems have been investigated. The classes most commonly compared have been between COVID-19, pneumonia, and normal images. In a study by Rahimzedah and Attar [52], they used two open-source datasets and utilized 180 X-ray images that were provided, where 118 cases were COVID-19 related were other lung infection-related diseases(SARS). Pneumonia cases had 6012 cases and Normal cases had 8851 normal cases. In this study, They trained their images on the ResNet50V2 and Xception among others, and also concatenated the models. The concatenated model achieved an average accuracy of 99.50% on the COVID-19 class and a sensitivity of 80.53%. In the study by Jain et al. [35], also another 3-class classification problem, the dataset collected had 6432 chest X-rays in total where COVID-19 only contributed 490 Images to the data. In this study, they implemented the LeakyReLU activation instead of the ReLU activation function. The model performed very well and obtained an accuracy of 97%. Other studies performing 3-class classification used different techniques to train on their data [35]. In the study by Ozturk et al. [72], where they used a dataset of 1127 CXR where COVID-19 contributes 127 Images. They used a technique called DarkNet and trained their data using a 5-fold cross-validation partitioning and obtained an accuracy of 98.08% and specificity and sensitivity both sitting at the 95% mark [72]. In another study by Luz et al.[32], they trained their EffetientNet model on a partition of 90% training set and 10]% testing set. With only 183 images of COVID-19 out of the 30000 cases, they achieved an accuracy of 93.9%, a sensitivity equal to 96.8%, and a precision rate of 100%.

In a study by Venkataramana et al. [49], they carried out a multiclass classification work where they looked into Tuberculosis and also divided Pneumonia into Bacterial and viral strains. They applied data augmentation to help reduce overfitting when training the network. They also used the SMOTE (Synthetic Minority Oversampling Technique) algorithm to balance the classes. Here they used 3500 cases of Tuberculosis and 1281 COVID-19 cases. Before applying SMOTE, they obtained an accuracy of 95.7% and 96.6% after balancing the data using SMOTE [49]. Shelke et al.[14], in their 4-class(COVID-19, Pneumonia, Tuberculosis(TB) and Normal) classification work, used various pre-trained models (DenseNet, VGG-16, ResNet-18). They created a one-to-many technique to reform binary classes to differentiate between two classes. VGG-16 achieved an accuracy of 95,9% segregating tuberculosis and pneumonia. To differentiate between COVID-19 and pneumonia, DenseNet-161 achieved a test accuracy of 98.9%.

In the current work, CoroNet outperformed ResNet-50 and VGG-19 and was presented for further retraining and testing on a 5-fold cross-validation partitioning on two datasets. CoroNet performed well on both of these datasets. The model obtained relatively high scores for accuracy, precision, specificity, and f1-score for the COVID-19 class with 82.00%, 89.18% 82.49%, and 81.88% respectively. However, sensitivity achieved a score of 71.13% on the augmented dataset. The results compared to other studies are relatively low and this may be to several reasons, which will be discussed throughout this chapter. The points to note in the study are as follows:

1) In this study, no manual feature selection, or extraction has been performed. The classification was directly performed end-to-end with raw data.
2) In this study, more data has been used compared to other studies in the literature.
3) This study contains the highest number of COVID-19 cases than other studies in the literature.
4) This study is one of the few studies in the literature that considers 5 classes in its study.
5) It has been studied and compared with transfer learning techniques.
6) The performance of COVID-19 data across normal, viral pneumonia and bacterial pneumonia, and tuberculosis was higher.

The observations made focused on a plethora of aspects. They were based on the matrices used to evaluate these outcomes, namely, accuracy, specificity and sensitivity, and other factors including how well they did use computational complexity as well as considering their architectural complexity of ResNet-50 and VGG-19. CoroNet architecture was less complex and had fewer parameters compared to ResNet-50. In previous work [10][71][47], ResNet50 and VGG19 achieved performance above 90% and this would be due to the dataset having fewer classes to help it learn and differentiate between. CoroNet also achieved a greater sensitivity compared to the other transfer learning-based architectures. This ensures that fewer false positives are being reported by this architecture, which is important since patients with COVID-19 require to be identified. The high false positive and false negative rate can lead to the spread and transfer of this deadly virus, which will lead to a significant burden on the healthcare system where resources are already exhausted and limited.

When explaining a model, seek to satisfy all the properties of XAI [68]. These include fairness, trust, usability, transparency, and causality [6]. Finding attribution methods that are simple and that can be applied to architectures that vary. Gary et al. [69], in their study, proposed the use of SmoothTaylor. They also used new measures like multi-scaled average total variations as new measures for noise for saliency maps. SmoothTaylor was able to show attribution maps that, compared to IG, are more relevance-sensitive and show less noise [68]. The current work was able to highlight pixels that contributed to the model misclassifying the images. These pixels were highlighted around the same regions or feature space of the images. This can be verified in Figure 5.17 and 5.16 and also explained in the results of why misclassification may have occurred. Hyperparameter tuning and removing the noise in the images may have contributed to the model being able to classify the images correctly and reach significant explanation maps [69]. Our experiment tried to show the localization of the regions that may have been mistaken as classes of one another and maybe a source of helping us understand the predictions made by the CoroNet model.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Medical Imaging had played a crucial role in combatting not only COVID-19 infection. The COVID-19 pandemic has left the world and individuals in a great depression. The shortage of resources and healthcare practitioners have brought a toll to the economic state and functionality of many countries. It is very important to be able to identify and isolate positive infections as quickly as possible to stop the spread of the infections. In this study, three CNN architectures were investigated to be able to classify COVID-19 and other related lung infections based on chest X-rays. Two models leveraged transfer learning, and the state-of-the-art architecture was built and designed to specifically identify COVID-19. About 22 000 CXR images from various sources were combined to allow the learning of the features that contribute to the decision-making. We also leveraged integrated gradients as our Explainable-AI method to visually see how and what features (pixels in this case) were used to make these predictions. while not a production-ready tool, more work and a clear understanding of these techniques may potentially make it become a powerful instrument in aiding clinicians and healthcare workers on the front line to identify positive patients with minimal contact with them by using computer-aided identification tools.

6.2 Future Work

Explainable-AI is a study that is sparking interest, especially in the field of medicine and healthcare. Looking at more techniques that show how these models do their predictions will bring more openness and trust to the use of these computer-aided methods. Trying to implement techniques like LRP, SmoothTaylor and Topological Data Analysis may contribute to better understanding, layer by layer, where misclassification may occur before reaching the output.

Bibliography

[1] AB.Nassif, I.Shahin, and I.Attili. “Speech recognition using deep neural networks: A systematic review”. In: *IEEE access* 7 (2019), pp. 19143–19165.

[2] A.Esteva, B.Kuprel, and RA.Novoa. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* 542.7639 (2017), pp. 115–118.

[3] AH.Al-Timemy, RN.Khushaba, and ZM.Mosa. “An efficient mixture of deep and machine learning models for covid-19 and tuberculosis detection using x-ray images in resource limited settings”. In: *Artificial Intelligence for COVID-19* (2021), pp. 77–100.

[4] A.Hedstr"om, L.Weber, and D.Krakowczyk. “Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond”. In: *Journal of Machine Learning Research* 24.34 (2023), pp. 1–11.

[5] A.Holzinger. “Explainable AI and multi-modal causability in medicine”. In: *i-com* 19.3 (2020), pp. 171–179.

[6] A.Holzinger, A.Saranti, and C.Molnar. “Explainable AI methods-a brief overview”. In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer. 2022, pp. 13–38.

[7] AI.Khan, JL.Shah, and MM.Bhat. “CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images”. In: *Computer methods and programs in biomedicine* 196 (2020), p. 105581.

[8] A.Jha, JK.Aicher, and MR.Gazzara. “Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study”. In: *Genome biology* 21.1 (2020), pp. 1–22.

[9] A.Kapishnikov, S.Venugopalan, and B.Avci. “Guided integrated gradients: An adaptive path method for removing noise”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5050–5058.

[10] A.Krizhevsky and G.Hinton. "Learning multiple layers of features from tiny images". In: (2009).

[11] A.Narini, C.Kaya, and Z.Pamuk. "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks". In: *arXiv preprint arXiv:2003.10849* (2020).

[12] A.Prasoon, K. Petersen, and C.Igel. "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2013, pp. 246–253.

[13] A.Sharma and S.Raniand D.Gupta. "Artificial intelligence-based classification of chest X-ray images into COVID-19 and other infectious diseases". In: *International journal of biomedical imaging* 2020 (2020).

[14] A.Shelke, M.Inamdar, and V.Shah. "Chest X-ray classification using deep learning for automated COVID-19 screening". In: *SN computer science* 2.4 (2021), pp. 1–9.

[15] A.Shrikumar, P.Greenside, and A.Kundaje. "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMLR. 2017, pp. 3145–3153.

[16] K.El Asnaoui, Khalid, and Y.Chawki. "Using X-ray images and deep learning for automated detection of coronavirus disease". In: *Journal of Biomolecular Structure and Dynamics* (2020), pp. 1–12.

[17] A.Wong, ZQ.Lin, and L.Wang. "COVID-Net S: Towards computer-aided severity assessment via training and validation of deep neural networks for geographic extent and opacity extent scoring of chest X-rays for SARS-CoV-2 lung disease severity". In: *arXiv preprint arXiv:2005.12855* (2020).

[18] J.Marco Blanco. "Detection of misclassified and adversarial examples in Deep Learning". PhD thesis. Universidad Politécnica de Madrid, 2019.

[19] John Hopkins Coronavirus Resource Centre. "An interactive web-based dashboard to track COVID-19 in real time". In: (). accessed 2021-03-29. URL: https://coronavirus.jhu.edu/map.html.

[20] CIFAR-10 dataset. "Kaggle CIFAR-10 Dataset. Retrieved May 2021". In: (). accessed 2021-07-16. URL: https://www.kaggle.com/c/cifar-10/data.

[21] CIFAR-100 dataset. "Kaggle CIFAR-100 Dataset. Retrieved May 2017". In: (). accessed 2021-07-16. URL: https://www.kaggle.com/c/ml2016-7-cifar-100.

[22] COVID-19 Chest X ray dataset. "Novel Coronavirus (2019-nCoV) Chest X-ray dataset. Retrieved April 2021". In: (). accessed 2021-04-28. URL: https://github.com/ieee8023/covid-chestxray-dataset.

[23] Kaggle Pneumonia dataset. "Pneumonia Dateset, Retrieved April 2021". In: (). accessed 2021-04-28. URL: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[24] Kaggle Tuberculosis dataset. "TB Dataset, Retrieved April 2021". In: (). accessed 2021-04-28. URL: https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities?select=ChinaSet_AllFiles.

[25] MNIST dataset. "Kaggle MNIST Dataset. Retrieved May 2021". In: (). accessed 2021-07-16. URL: https://www.kaggle.com/oddrationale/mnist-in-csv.

[26] D.Hendrycks and K.Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[27] DS.Kermany, M.Goldbaum, and W.Cai. "Identifying medical diagnoses and treatable diseases by image-based deep learning". In: *cell* 172.5 (2018), pp. 1122–1131.

[28] D.Wang, J.Mo, and G.Zhou. "An efficient mixture of deep and machine learning models for COVID-19 diagnosis in chest X-ray images". In: *PloS one* 15.11 (2020), e0242535.

[29] E.Amann, A.Blasimme, and E.Vayena. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC Medical Informatics and Decision Making* 20.1 (2020), pp. 1–9.

[30] E.Dong, H.Du, and L.Gardner. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet infectious diseases* 20.5 (2020), pp. 533–534.

[31] EE.Hemdan, MA.Shouman, and ME.Karar. "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images". In: *arXiv preprint arXiv:2003.11055* (2020).

[32] E.Luz, P.Silva, and R.Silva. “Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images”. In: *Research on Biomedical Engineering* 38.1 (2022), pp. 149–162.

[33] FV.Farahani, K.Fiok, and B.Lahijanian. “Explainable AI: A review of applications to neuroimaging data”. In: *Frontiers in Neuroscience* 16 (2022).

[34] G.Bebis, D.Egbert, and M.Shah. “Review of computer vision education”. In: *IEEE Transactions on Education* 46.1 (2003), pp. 2–21.

[35] G.Jain, D.Mittal, and D.Thakur. “A deep learning approach to detect Covid-19 coronavirus with X-ray images”. In: *Biocybernetics and biomedical engineering* 40.4 (2020), pp. 1391–1405.

[36] GSW.Goh, S.Lapuschkin, and L.Weber. “Understanding integrated gradients with smoothtaylor for deep neural network attribution”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 4949–4956.

[37] H.Aboutalebi, M.Pavlova, and MJ.Shafiee. “COVID-Net CXR-S: Deep Convolutional Neural Network for Severity Assessment of COVID-19 Cases from Chest X-ray Images”. In: *arXiv preprint arXiv:2105.00256* (2021).

[38] H.Panwar, PK.Gupta, and MK.Siddiqui. “Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet”. In: *Chaos, Solitons & Fractals* 138 (2020), p. 109944.

[39] I.Castiglioni, D.Ippolito, and M.Interlenghi. “Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: a first experience from Lombardy, Italy”. In: *European Radiology Experimental* 5.1 (2021), pp. 1–10.

[40] ID.Apostolopoulos and TA.Mpesiana. “Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks”. In: *Physical and Engineering Sciences in Medicine* 43.2 (2020), pp. 635–640.

[41] I.Guyon and A.Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[42] Normal Chest X ray images. “Chest X-ray 8. Retrieved April 2021”. In: (). accessed 2021-04-28. URL: https://github.com/TRKuan/cxr8.

[43] J.Hirschberg and CD.Manning. “Advances in natural language processing”. In: *Science* 349.6245 (2015), pp. 261–266.

[44] J.Kim, A.Rohrbach, and Z.Akata. "Toward explainable and advisable model for self-driving cars". In: *Applied AI Letters* (2021), e56.

[45] JRA.Solares, FED.Raimondi, and Y.Zhu. "Deep learning for electronic health records: A comparative review of multiple deep neural architectures". In: *Journal of Biomedical Informatics* 101 (2020), p. 103337.

[46] K.Simonyan, A.Vedaldi, and A.Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[47] L.Brunese, F.Mercaldo, and A.Reginelli. "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays". In: *Computer Methods and Programs in Biomedicine* 196 (2020), p. 105608.

[48] L.Hussain, T.Nguyen, and H.Li. "Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection". In: *BioMedical Engineering OnLine* 19.1 (2020), pp. 1–18.

[49] L.Venkataramana, D.Prasad, and S.Saraswathi. "Classification of COVID-19 from tuberculosis and pneumonia using deep learning techniques". In: *Medical & Biological Engineering & Computing* 60.9 (2022), pp. 2681–2691.

[50] L.Wang, ZQ.Lin, and A.Wong. "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images". In: *Scientific Reports* 10.1 (2020), pp. 1–12.

[51] M.Pavlova, N.Terhljan, and AG.Chung. "COVID-Net CXR-2: An Enhanced Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-ray Images". In: *arXiv preprint arXiv:2105.06640* (2021).

[52] M.Rahimzadeh and A.Attar. "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2". In: *Informatics in medicine unlocked* 19 (2020), p. 100360.

[53] N.Buch, SA.Velastin, and J.Orwell. "A review of computer vision techniques for the analysis of urban traffic". In: *IEEE Transactions on intelligent transportation systems* 12.3 (2011), pp. 920–939.

[54] N.Tsiknakis, E.Trivizakis, and EE.Vassalou. “Interpretable artificial intelligence framework for COVID-19 screening on chest X-rays”. In: *Experimental and Therapeutic Medicine* 20.2 (2020), pp. 727–735.

[55] N.Ueda. “Optimal linear combination of neural networks for improving classification performance”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.2 (2000), pp. 207–215.

[56] World Health Organization. “Novel Coronavirus (2019-nCoV) Situation Report–28. Retrieved March 2020”. In: (). accessed 2021-03-28. URL: https://www.who.int/docs/defaultsource/coronaviruse/situation-reports/20200217-sitrep-28-covid-19.pdf?sfvrsn=a19cf2ad_2.

[57] World Health Organization. “Taxonomy of Virus”. In: (). accessed 2021-03-28. URL: http:www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it#:~:text=Viruses%20are%20named%20based%20on,Viruses%20(ICTV)..

[58] O.Russakovsky, J.Deng, and H.Su. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[59] PK.Sethy, SK.Behera, and PK.Ratha. “Detection of coronavirus disease (COVID-19) based on deep features and support vector machine”. In: (2020).

[60] P.Linardatos, V.Papastefanopoulos, and S.Kotsiantis. “Explainable ai: A review of machine learning interpretability methods”. In: *Entropy* 23.1 (2020), p. 18.

[61] R.Jain, M.Gupta, and S.Taneja. “Deep learning based detection and analysis of COVID-19 on chest X-ray images”. In: *Applied Intelligence* (2020), pp. 1–11.

[62] R.Miotto, F.Wang, and S.Wang. “Deep learning for healthcare: review, opportunities and challenges”. In: *Briefings in bioinformatics* 19.6 (2018), pp. 1236–1246.

[63] R.Miotto, F.Wang, and S.Wang. “Deep learning for healthcare: review, opportunities and challenges”. In: *Briefings in bioinformatics* 19.6 (2018), pp. 1236–1246.

[64] S.Gu, M.Pednekar, and R.Slater. “Improve Image Classification Using Data Augmentation and Neural Networks”. In: *SMU Data Science Review* 2.2 (2019), p. 1.

[65] SH.Yoo, H.Geng, and TL.Chiu. “Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging”. In: *Frontiers in medicine* 7 (2020), p. 427.

[66] S.Khobahi, C.Agarwal, and M.Soltanalian. “Coronet: A deep network architecture for semi-supervised task-based identification of covid-19 from chest x-ray images”. In: *MedRxiv* (2020).

[67] S.Khobahi, C.Agarwal, and M.Soltanalian. “Coronet: A deep network architecture for semi-supervised task-based identification of covid-19 from chest x-ray images”. In: *MedRxiv* (2020).

[68] S.Sattarzadeh, M.Sudhakar, and KN.Plataniotis. “Integrated grad-CAM: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 1775–1779.

[69] SWG.Goh, S.Lapuschkin, and L.Weber. “Understanding integrated gradients with smoothtaylor for deep neural network attribution”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 4949–4956.

[70] S.Xu, J.Wang, and W.Shou. “Computer vision techniques in construction: a critical review”. In: *Archives of Computational Methods in Engineering* 28 (2021), pp. 3383–3397.

[71] T.Ozturk, M.Talo, and EA.Yildirim. “Automated detection of COVID-19 cases using deep neural networks with X-ray images”. In: *Computers in biology and medicine* 121 (2020), p. 103792.

[72] T.Ozturk, M.Talo, and EA.Yildirim. “Automated detection of COVID-19 cases using deep neural networks with X-ray images”. In: *Computers in biology and medicine* 121 (2020), p. 103792.

[73] V.Gulshan, L.Peng, and M.Coram. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *Jama* 316.22 (2016), pp. 2402–2410.

[74] Y.Kene, U.Khot, and I.Rizvi. “A Survey of Image Classification and Techniques for Improving Classification Performance”. In: *Proceedings 2019: Conference on Technologies for Future Cities (CTFC)*. 2018.

[75] Y.LeCun, Y.Bengio, and G.Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.