Detection of COVID-19 Infection with an Explainable, Hybrid CT/X-Ray Model

Student Name: Thomas Pettit

Supervisor Name: Yang Long

Submitted as part of the degree of MEng Computer Science to the

Board of Examiners in the Department of Computer Sciences, Durham University

Abstract — The COVID-19 pandemic has shown the importance of rapid low-cost diagnosis of COVID-19 cases. By accelerating the process of diagnosing a patient with COVID-19, doctors and other medical personnel have more time available to see other patients, instead of analysing these scans, which also results in saving medical institutions money. Thus, this project aims to bring advances in computer vision and machine learning to the domain of COVID-19 infection classification in order to achieve this. In this project, chest X-Ray and CT scans associated with COVID-19 cases, healthy individuals, and other respiratory diseases such as pneumonia are combined to form a hybrid dataset to be used to train 5 machine learning models for the task of classifying between examined classes. By creating a hybrid dataset, it removes the need for two different models to be trained used by medical personnel and institutions, saving money and time. Due to the importance of explanations of these predictions in the medical field in order to build trust, explainability in the form of heatmaps are implemented for each model in order to identify key areas in the image used by these models in order to make their predictions. Then, a new ensemble model is proposed, which brings both performance improvements as well as significantly improved explainability across various metrics. The performance of these 5 models is then compared using various evaluation metrics, and the most effective model is found to be the proposed model. Additionally, many respiratory diseases share common characteristics in the way that they can be diagnosed from medical images, and as a result the findings of this paper can easily be used to help automate the diagnosis of other respiratory illnesses.

Index Terms—Computer Vision, Image Processing, Machine Learning, CNNs

1 Introduction

As computer vision and machine learning have become more powerful, this project seeks to examine how novel machine learning models and image analysis techniques can be applied to find the most effective model capable of detecting COVID-19 infection from a hybrid dataset of both X-Ray scans and CT scans, whilst also providing explanations for its decision. These advances are then built upon by a proposed ensemble model, which seeks to improve upon the state-of-the-art.

1.1 Background

In 2019, a novel coronavirus disease was discovered in Wuhan, China, named COVID-19 by the World Health Organisation (WHO) [1]. COVID-19 is an extremely infectious disease, which has been declared a pandemic by the WHO [2]. The virus has managed to spread across the world, despite Governments of different countries imposing restrictions such as social distancing, border restrictions and increasing awareness of hygiene. Most people who contract the disease experience mild to moderate symptoms, but some develop deadly pneumonia. Many techniques were used to try and detect a COVID-19 infection, such as measuring body temperature, reverse-transcription-polymerase chain reaction (RT-PCR), CT scans and X-Ray scans. However, there have been issues identified for some of these methods. For example, measuring body temperature using infrared thermometers or thermal image scanners was found to not be sufficiently accurate, often poorly standardized and not effective [3]. The RT-PCR test can also lead to false negatives. Two studies found that between 3% and 30% of COVID-19 patients who initially had a negative RT-PCR test, then showed a positive chest CT scan for COVID-19 a few days later, which was then later confirmed by a second RT-PCR [4], [5]. One study, comparing the sensitivity of chest CT scans for COVID-19 to RT-PCR tests [5], found that in a sample of 51 patients who were positive for COVID-19, CT scans detected the positive infection in 50 of 51 cases, whilst RT-PCR only detected a positive infection in 36 of 51 patients. This therefore shows that the low sensitivity of RT-PCR tests, also supported in other literature such as [6], means that other reliable methods to screen COVID-19 patients are required. Chest CT scans and X-Rays are a non-invasive method of achieving this, with their use for this purpose supported in literature [7], [8]. However, despite their suitability, medical personnel often struggle to detect small changes in these scans caused by the COVID-19 disease, because COVID-19 can display similar characteristics to other respiratory diseases in medical images. As a result, machine learning techniques are required to improve both accuracy and speed of detection, and aid medical personnel in their diagnosis from these medical images.

CT scans contain features that can describe characteristics of infected tissue [9], and can therefore be used to detect COVID-19 infection. Many studies, [10], [11], have investigated the suitability of these features in CT scans for distinguishing between COVID-19 infection and other

respiratory infections. Unfortunately, COVID-19 produces CT scan features that are similar to those caused by pneumonia [12]. Additionally, one study [13] has shown that COVID-19 can mimic disease processes of other diseases, including other infections, which can also lead to a misdiagnosis of COVID-19 with other pneumonia. As a result, intelligent tools are required to help detect these very slight changes in features that can help to diagnose a positive COVID-19 CT scan.

In order to build these intelligent tools, novel advances in machine learning can be used. Machine learning techniques have recently been shown to perform very well analyzing complex medical data. Deep learning algorithms such as the Convolutional Neural Network (CNN) have been shown to perform particularly well when automatically processing large amounts of medical images, and identifying complex associations in high-dimensional data for diagnosing diseases, as shown in [14]. Recently, deep learning models have been successfully applied on CT [15] and X-Ray [16] scans for the automated detection of COVID-19.

However, despite recent work [17] showing deep learning CNN models are successful for predicting medical outcomes compared to traditional radiomic pipelines, the implementation of this strategy is prone to overfitting. This is typically due to the limited datasets available with labelled images. Previous studies have implemented transfer learning, where convolutional features learned from a related image processing task can be reused to advance the learning of a new image processing task, which reduces the need for a significantly larger dataset.

As a result, there is a clear requirement for automatic and accurate detection of COVID-19 infection. Temperature checks and RT-PCR tests have been shown to be unsuitable, and chest X-Rays and CT scans have had many studies backing their suitability for this task, but human diagnosis from these medical images has been found to be flawed due to potential similarities between COVID-19 infection and other respiratory diseases. As a result, this project will focus utilizing both X-Ray and CT scans, as well as novel machine learning models from the relevant literature, in order to solve this task.

1.2 Motivation

By accelerating the process of detecting COVID-19 from a CT scan or X-Ray scan, it means that doctors and other medical personnel will have more time available to see other patients, instead of having to intricately analyse these scans to predict whether the patient has a COVID-19 infection. This would save medical institutions money, as they are able to see more patients per day with the same number of doctors.

Additionally, with the ending of free mass symptomatic and asymtomatic testing for the general public for COVID-19 in the UK from April 1st 2022, [18], it has become difficult for members of the public to know any long-term consequences caused by a COVID-19 infection in the future. This is because the requirement to pay for a test to detect COVID-19 has meant some people are unable to afford them, and it will act as a deterrent for other people who may have symptoms of COVID-19 but decide it is not worth them paying to find out if they have contracted the disease or not. This means that if these members of the public develop any illnesses or diseases in future, they will be unaware whether this is a long-term consequence of a past COVID-19 infection, or another illness they should investigate. This problem has been worsened by the recent cost of living crisis in the United Kingdom in 2022 and 2023.

By looking at CT scans or X-Rays of a patient, and automatically detecting a COVID-19 infection using a deep learning model, medical personnel can both improve the accuracy of their diagnosis and the speed at which they can make it. This means that members of the public will then be aware that they have had or currently have the disease more quickly, and also with greater confidence. As a result, patients will know whether they are more likely to have long-term effects from the disease, such as scarring of the lung tissue. This would inform the patient to be more careful about contracting other respiratory illnesses, such as pneumonia, in the future, as it could be potentially more dangerous for them.

Another motivation for this project is due to the fact that many respiratory diseases share common characteristics in the way that they can be diagnosed from medical images. This means the findings in this project can be easily used to help automate the diagnosis of other respiratory illnesses, for example through the use of transfer learning, and so this research can be potentially used to automate detection of future diseases similar to COVID-19.

1.3 Objectives

The research question that this project will address is: *What are the most effective machine learning models that can make explainable predictions on whether a CT or X-Ray scan shows evidence of a positive COVID-19 infection?* This will be achieved by building upon CNN architectures implemented in other literature, namely: VGG16 [19], DenseNet201 [20] , DarkNet19 [21] and EfficientNetB0 [22], as these models have been shown to produce high levels of accuracy, as well as performed well with the evaluation metrics used in each paper in similar research areas. A model that can operate on both X-Ray and CT scans means a separate model is not required for both types of images. The ability for a single model to be able to operate on both types of images has also been shown in other literature, with [23] confirming that CNNs using transfer learning can predict COVID-19 in both chest X-Ray and CT scans individually. Then, as most of these models do not have an explainable element to their predictions, this will be built on top of these models. This explainable element is important in this domain because, according to an international statement on the ethics of artificial intelligence in radiology, "transparency, interpretability, and explainability are necessary to build patient and provider trust". [24] Then, a new ensemble model, built on top of these state-of-the-art model architectures is proposed, which yields increased performance on evaluation metrics and improved explainability.

There are three elements to this project that make it novel. First is to produce a machine learning model that works on both X-Ray and CT scans (a hybrid model). Second is to produce a hybrid model that generates explainable predictions as to why a scan does or does not indicate a positive COVID-19 infection. The combination of these two elements has not been done to date in the literature.

Then, a new ensemble model is proposed which yields improved performance on both evaluation metrics and explainability, compared to the current state-of-the-art.

2 RELATED WORK

The work in this paper builds upon advances in image processing techniques, machine learning for classification problems and explainable predictions of classifier models. Most previous attempts at predicting whether CT or X-Ray scans show evidence of a COVID-19 infection have built models for either X-Ray or CT scans, but not both, such as [15], [26] and [27]. These models have been shown to produce good results in terms of accuracy and other metrics, mainly due to the advances in computer vision and machine learning, such as CNNs, but lack the ability to operate on both types of images, as well as generate explanations for their predictions. In order to provide an overview of the techniques and methodologies of these related works, 7 different factors can be investigated.

A. Datasets

The dataset different approaches have used is an important issue. Many approaches in related works have used datasets with 7500 datapoints or less. As mentioned previously, most approaches have just used one type of medical image, either CT or X-Ray scans, which is one of the causes for this. This limited dataset size has lead to many implementations harnessing the power of transfer learning in order to train their machine learning models. This is because the transfer learning technique allows for quick retraining of very deep CNNs with a comparatively low number of images. For example, the concept was used by Vikash et al [28] to detect pneumonia using a pretrained ImageNet model and Xianghong et al [29] implemented a custom VGG16 model for identifying lung regions and different types of pneumonia classification. The fact that these implementations have used these smaller datasets means that is is difficult to generalize their results, and it is not guaranteed that the performance achieved by these papers will hold when the models are tested on a larger dataset. As a result, investigation into how these models perform when trained upon a larger dataset needs to be done. This is part of the research this project aims to fulfill.

However, the creation of a larger dataset for this brings rise to several difficulties. Many papers, such as [5], [16], [27] and [30] to name a few, were written in 2020. Thus, at the time the papers were written, there was a lack of availability of public images of COVID-19 patients. Many of these papers then had to use certified radiologists to manually label the images for use in machine learning models, a time consuming process. Additionally, many papers have attempted to compile the medical images from various sources, in order to build their own dataset. As a result, many papers have created their datasets with portions of their data from the same source, for example the covid-chestxray-dataset [31], which at the time of writing has been cited by 677 papers. This means that simply appending one dataset onto another most likely gives rise to duplicate images in the new dataset. If this dataset forms a large majority of the entire dataset used by particular papers, then combining these datasets would mean the majority of datapoints in the new, larger dataset would be duplicates, invalidating this study on larger datasets. Moreover, the datasets often collect data from public sources and hospitals and physicians, meaning the datasets used by different papers may not be the same, but could still contain the same data if they happen to have been collected from the same initial source. By combining both X-Ray and CT scans, a larger dataset can be created, whilst minimizing the chance of duplicates in the dataset. This is because it is unlikely that there are both publicly available CT and X-Ray scans for the same person taken at the same time. This approach was followed by a few papers, such as [23], which combined 846 CT scans and 657 X-Ray scans, but still this dataset still only contained approximately 1500 samples. This approach can thus be beneficial in the creation of a larger dataset in this project.

Due to the fact that these datasets are using health data from individuals in the form of CT or X-Ray scans of a person's chest, ethics are an important issue to consider when collecting data for the datasets. As stated previously, many papers use the same datasets in their implementations, with most of these datasets being approved by ethics committees, or clearly stating how the data was ethically collected. For example, the covid-chest-xray dataset [31] was approved by the University of Montreal's Ethics Committee and the data collected in the study by Chaddad et al. [23], clearly states their data collection procedure.

B. 2 Class vs 3+ Class Problem

This classification task of identifying a COVID-19 infection from CT or X-Ray scans can be achieved via different methods. The dataset must be compromised of both scans that are positive for a COVID-19 infection, and scans that are of people who do not have a COVID-19 infection. Different papers approached this in different ways. For the non-COVID-19 scans, some papers selected scans of healthy individuals, with no other respiratory diseases, such as [32], and thus only used positive COVID-19 scans and healthy lungs in their dataset. This is known as the 2 class problem, as there are only two possible classifications for each datapoint. Other papers opted for a 3 class problem. This is where the dataset includes positive COVID-19 scans, healthy individuals, and one other respiratory disease, such as pneumonia in [32]. Some papers took this a step further, and included in their dataset more respiratory illnesses, such as the paper by Minaee et al. [30], which included 13 other respiratory illnesses in their dataset, such as Edema. A dataset containing more than one type of scans in the 'other' category, such as the study by Minaee et al., is known as the three or more class problem. The three or more class approach can bring performance benefits for the classifiers, due to the fact that features learnt from the images for the COVID-19 positive class by a machine learning model in the two class problem may actually be more indicative of other respiratory illness, rather than COVID-19, such as pneumonia. However, due to the fact that the various other respiratory illnesses in 'other' category of the dataset may share very similar characteristics to COVID-19, this could lead to the classifier struggling to differentiate between these similar looking diseases, particularly if the dataset contains fewer samples to help the models learn the best method of classification.

C. Image Augmentation

Due to the limited dataset size, most papers used image augmentation to increase the number of datapoints available to train their machine learning models. This is often done via different techniques such as flipping the images, rotating the images slightly, adding a small amount of distortions to the existing images or translating the images horizontally or vertically. This therefore artificially creates new scans to expand the dataset, with many papers reporting a sizeable increase in their dataset size. For example, [30] created a five-fold increase in the number of samples. The theory behind this approach is that, with more training data, the machine learning models can become better at generalizing, and so reduce the chance of models overfitting to the training data, as well as become more accurate at the classifying task. However, some studies have found this difference to be marginal, for example [32], which found that its best performing model in the two-class problem only performed very slightly better when trained with augmented images compared to without, with 99.41% accuracy without, and 99.69% accuracy with augmented images. On the other hand, the performance differences vary depending on whether the paper is investigating a two-class or three or more class problem. Study [32] also found that, in the three class problem, their DenseNet model's accuracy increased from 95.19% without image augmentation, to 97.94% with image augmentation, as well as large improvements in other evaluation metrics such as precision and sensitivity. Other studies have also backed the effectiveness of data augmentation for machine learning models, such as the study by Perez et al. [33].

D. Models

Different papers in similar research areas implemented different machine learning models in order to solve their respective classification task. However, the majority of the studies created a CNN and used transfer learning alongside image augmentation, such as [30] and [32]. This is because transfer learning with CNNs is well-suited to detecting anomalies in medical images, as these anomalies are normally characterized by local changes in texture instead of high-level structures in the image.

Different papers have either selected one model and tried to optimize its performance, or have trained several different models and compared their performance. As stated previously, transfer learning was commonly used. Das et al. [34] followed a deep-transfer learning approach, implementing a modified Inception model. Gianchandani et al. [35] proposed two ensemble deep transfer learning models for the two-class problem. The proposed ensemble model's (a combination of VGG16 and DenseNet) performance was compared with VGG16, ResNet152V2, InceptionResNetV2 and DenseNet201, and was found to be the best performing, with 96.15% accuracy in the two-class problem. However, deep learning approaches were also used by several papers, without harnessing the power of transfer learning. For example, Alakus et al. [36] implemented 4 different types of deep learning models, and 2 hybrid models. They found that the highest accuracy was found by the CNNLSTM hybrid model, with 92.3%.

Many studies, such as [17], have shown how deep CNNs are advantageous in predicting clinical outcomes compared to traditional radiomic pipelines. However, this approach is prone to overfitting due to the small dataset sizes. Thus, alongside image augmentation, other novel techniques have been suggested, for example the study by Chaddad et al. [37], which suggests entropy-related features extracted at different layers of a CNN to train a separate classifier model for the final prediction.

Despite the large variation in implemented models for this classification task, the conclusion is still the same. Either pure deep learning models, or models harnessing transfer learning, can be used to achieve high performances in various evaluation metrics, and this is clearly supported in the relevant literature.

E. Hybrid Models

The application of a single model to a dataset containing both CT and X-Ray scans is uncommon. Most studies trained either a single model, or compared the performance of various different models, upon a dataset consisting of only one type of medical image, such as [15] and [16]. However, some studies trained their models on both types of images, for example the study by Chaddad et al. [23], trains 6 well-known CNN architectures on a dataset containing 846 CT images and 657 X-Ray images. The study found that, by combining both types of images, the DarkNet architecture achieved the highest accuracy with 99.09% in the two-class problem. This therefore showed that a single model could be used to predict positive COVID-19 infections from both types of images.

As stated previously, by using both types of data, this can offset the issues found by many studies of the lack of available scans of positive COVID-19 patients. This, coupled with image augmentation, can create datasets of sufficient size to train the novel deep learning models suggested in some of the studies.

F. Explainable Predictions

Most of the studies aiming to classify a positive COVID-19 infection from either CT or X-Ray scans or both do not generate explanations for their decisions. The issue is that deep CNNs are not inherently interpretable, and are often seen as black boxes. As stated previously, by the international statement on the ethics of AI in radiology, "transparency, interpretability, and explainability are necessary to build patient and provider trust" [24]. Clinicians would likely not put their trust in the output of an algorithm if it did not explain its reasoning behind its prediction. From an ethical standpoint, it is ethically responsible to create explainable machine learning models. It also aids in detecting any unintended bias that may be created in machine learning models.

One particular study that did implement explainable predictions for positive COVID-19 scans on X-Ray images was the study by Blake VanBerlo and Matt Ross [38], who used the LIME algorithm. Another study that also generated explainable predictions was by Karim et al. [26]. This study generated class-discriminating attention maps using gradient-guided class activation maps (Grad-CAM++), outlined in [39], and layer-wise relevance propagation (LRP), outlined in [40], to provide explanations of the predictions, by identifying critical regions on the patient's chest. This helps to remove the opaqueness often

associated with black-box models via providing these human-friendly explanations for the predictions, as outlined in [41]. The concept of generating heatmaps for indicating the regions of interest in the input image is the most common use of explainable AI for COVID-19 prediction. This is to ensure that the model is focusing on the correct regions of interest (ROIs) in the image that are typically indicative of the presence of the COVID-19 disease. This approach has been followed by various other studies, such as those by Mei et al. [42], Bai et al. [43], Wehbe [44] and Murphy [45]. Interestingly, Murphy [45] and Wehbe [44] showed heatmaps for both positive and negative COVID-19 scans, and found that negative scans show low influence within the lungs. Alternatively, Bai et al. [43] generated heatmaps that also showed correct highlighting of areas showing COVID-19 disease within lung segmentations, however also found that regions with no content such as areas outside the lung mask were influential to the classification output. Similarly, Jin et al. [46] implemented both Grad-CAM and Guided Grad-CAM to visualize influential image regions. Similar to the work by Bai et al. [43], they found that Grad-CAM indicated that their model identified regions both inside and outside of the lungs as highly influential. However, they found that Guided Grad-CAM, whilst improving the heatmap visualisations, did not capture all the disease tissue. This would suggest that different variations of Grad-CAM can lead to variations in performance in terms of the explainability of the models.

G. Evaluation Metrics

Several different evaluation metrics are used across the literature. Typically, accuracy, sensitivity, specificity, ROC curve, AOC and confusion matrices are used, such as in the study by Minaee et al. [30]. Other metrics such as precision and F1-score are also included in other studies, such as the study by Ozturk et al. [27]. Another metric that aids in the discussion of the most effective model for making explainable predictions for whether a scan shows signs of a positive COVID-19 infection is the processing time. The processing time is the time taken for the model to make a prediction upon new data, once it has already been trained. This is because if a model takes too long to make its prediction, then it could lead to issues if a doctor has to wait a long period of time before they are presented with a prediction by the system. Thus, some studies have also included this metric in their work, such as the study by Li et al. [15]. However, this study found the average processing time for each CT examination by their model to be only 4.51 seconds, and so slight variations in this number depending on the type of model used are unlikely to be sufficiently large to create this problem.

3 METHODOLOGY

A. Data Preparation

In order to create a larger dataset containing both X-Ray and CT scans, enough data had to be found. In this paper, the three-plus-class problem is investigated. This is because, and as stated previously, the three or more class approach can often bring performance benefits for the classifiers when used in live inference, due to the fact that features learnt from the images by a machine learning model in the two-class problem may actually be more indicative of other respiratory illness, rather than COVID-19. Additionally, in reality if these machine learning models were used to assist medical personnel in their diagnosis of COVID-19 from medical images, a patient could have one of many different respiratory diseases. As such, the 3-class problem or the 2-class problem would not be a realistic representation of what the model would have to be able to run upon for inference. One of the main motivations for this project is also to let patients know if they have had a COVID-19 infection recently, or currently have a COVID-19 infection, so that they know whether they need to be more careful about contracting respiratory illnesses in future due to the long-term effects that COVID-19 can have. The three-plus-class problem will help in this case, due to the performance benefits that this approach can brings, as stated above.

The dataset itself is comprised of both X-Ray and CT scans. As the three-plus-class problem has been selected, two of the classes are COVID-19 positive, and healthy patient scans. Due to the abundance of chest X-Ray and CT scans showing positive cases for pneumonia and other respiratory diseases, pneumonia-positive and other disease scans were selected for the other classes. For the COVID-19 positive X-Ray scans, the covid-chestxray-dataset [31] was selected. This is because it is a very popular dataset amongst the literature, cited by 677 other papers at the time of writing. This dataset is also approved by the University of Montreal's Ethics Committee, and contains 468 images of COVID-19 positive scans. Only the COVID-19 scans will be taken from this dataset. In order to create a larger dataset including more COVID-19 positive X-Ray scans, image augmentation will be applied to these scans at random (discussed later). As this has been proven to create up to a five-fold increase in data quantity [30], this will create 2500 COVID-19 positive X-Ray scans. One issue with chest X-Ray or CT scan datasets is the fact that the data has to be labelled manually, often by an experienced radiologist. However, this labelling process may not always be 100% accurate, and so could lead to mistraining of the machine learning models. Thus, the datasets collected should contain images that are labelled to a high degree of confidence. Thus, for the healthy and other classes, the National Institute of Health's Chest X-Ray dataset [47], known as ChestX-ray8, was used. The labels for the images in this dataset are expected to be more than 90% accurate, and so are suitable for this study. This dataset contains scans that show patients diagnosed with other illnesses, such as Hernia, Fibrosis, Edema and Emphysema to name a few. Those labelled as "No Finding", represent the healthy class, and are selected to be the healthy scans. To ensure that all the classes in the study are roughly even sized, so as to prevent any imbalances, 2500 healthy scans will be taken from this dataset. For the other classes for the dataset, the ChestX-ray8 dataset will be used again, which contains 1062 images of pneumonia-positive chest X-Rays, as well as thousands of images of other diseases, as mentioned above. Thus, these other disease scans are selected at random from this dataset, and image augmentation applied. Again to ensure all classes are roughly even sizes, 2500 of these were selected at random.

For the COVID-19 positive CT scans, the SARS-CoV-2 CT-scan dataset [48] was used, as it is a popular dataset in

literature, and cited by 187 other papers at the time of writing. This dataset contains 1252 COVID-19 positive scans. Image augmentation (see below) was also be applied randomly to these scans, in order to create 2500 total COVID-19 positive CT scans. For the healthy CT scans, the COVID-CT dataset [49] was used. Again, this was because it is a popular dataset in the literature, with 367 citations, but also because, and as supported by Aswathy et al. [50], the non-COVID CT images from this dataset are complex, and it is a challenging task to distinguish either COVID-19 or non-COVID-19. Thus, when the models were trained on this data, they should be of higher quality, as the task was more difficult. The dataset was also confirmed by a senior radiologist in Tongji Hospital, Wuhan, China, who has performed diagnosis and treatment of a large number of COVID-19 patients during the outbreak of the virus. The dataset contains 463 non-COVID-19 CT images, and so all of these images were used. These images were augmented in order to create 2500 of these complex images for the healthy class. For the other classes, the SARS-Cov-2 CT-scan dataset was used again, as it contains 1230 scans of patients diagnosed with other pulmonary diseases. Image augmentation was also applied to these non-COVID-19 patient scans, in order to create 2500 of these scans.
Thus, the structure of the whole dataset in this study can be seen in Figure 1 below:

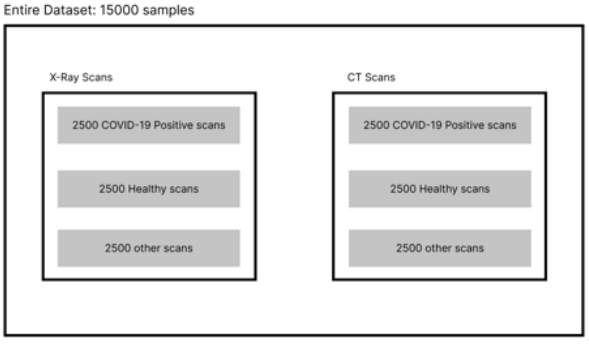

Figure 1: Dataset

All the images in the dataset will be resized to 256 by 256 pixels in order to be fed into the machine learning models. The images were also all converted to grayscale, and normalisation was applied to ensure all the images had a mean and standard deviation of 0.5.

Image Augmentation
In order to create a large increase in the size of the dataset, image augmentation had to be applied. This was achieved through the following techniques: horizontal flipping of the image, vertical flipping of the image, rotating of the image, translations of the image and perspective changes, as well as combinations of these transforms. By applying this range of transforms for image augmentation, the created dataset is of higher complexity for the machine learning models. This allows the models to generalise better to scans outside of their training data. Additionally, medical images can have issues with quality assurance, and some scans may have incorrect contrast or contain noise or artifacts on their images, as discussed in the study by Tompe et al. [51]. Thus, by applying these transforms to the images in the dataset, we can better replicate real-world examples of scans that may have to be fed through the models for inference. An example of the transformed images for the X-Ray scans can be seen in Figure 2 below:

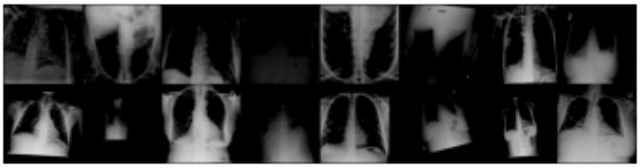

Figure 2: Example Batch from Dataloader

B. Model Selections

In this study, 4 different state-of-the-art machine learning models were selected to be compared against each other. These models were selected based upon their performances in their previous implementations in other literature. Certain popular models were excluded from this study, such as ResNet (specifically ResNet152V2) and Inception (specifically InceptionRestNetV2), due to their relatively low performance for the three-class problem for COVID-19 classification, as reported in the literature. Thus, the 4 chosen model architectures were: VGG, DenseNet, DarkNet, EfficientNet.

Specifically, the version of VGG implemented was VGG16 [19]. This is because Xianghong et al [29] implemented a custom VGG16 model for identifying lung regions and different types of pneumonia classification, and achieved an accuracy of 80.48%. Gianchandani et al. [35] found the VGG16 model in the task of diagnosing COVID-19 cases from chest radiographic images, achieved an accuracy of 95.7%. This high accuracy is critical to this study, and so this is one of the reasons this model was selected. The architecture for the VGG16 model is shown in Figure 3 below.

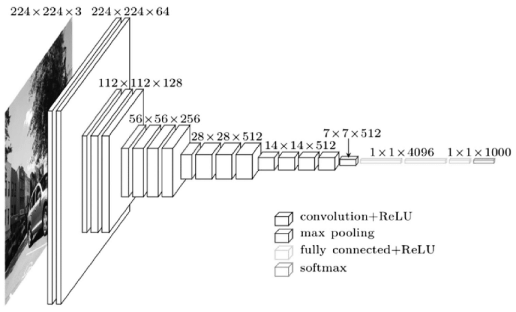

Figure 3: VGG16 Architecture [52]

In the VGG16 architecture, the input image was first passed through a stack of convolutional layers, where filters with a small receptive field (3x3) are used. The convolutional stride and the spatial padding of the convolutional layer input was fixed to 1 pixel for 3x3 convolutional layers. This was to ensure that the spatial resolution was preserved after convolution. Max pooling layers are then added after some of the convolution layers in order to aid in spatial pooling. This was performed with a 2x2 pixel window, with a stride of 2. There are then three fully connected layers that follow the stack of convolutional layers. The final layer is then the softmax layer. The architecture followed changes this architecture shown in

Figure 3 slightly, by modifying the final section to be an output of 3 dimensions (one for each class), rather than 1000. This is also followed for all the other 3 state-of-the-art models in this study.

The version of DenseNet implemented in this study was DenseNet201 [20]. This is because it also achieves a high accuracy in diagnosing COVID-19 from medical images. Gianchandani et al. [35] found that their implementation of DenseNet201 achieved an overall accuracy of 96.68% in the three-class problem. Similarly to this study, Gianchandani et al's three-class approach used COVID-19, normal and pneumonia as the three classes, where DenseNet201 achieved an F1 score of 0.98 for the COVID-19 and normal classes, and 0.95 for the pneumonia class. The study by Chowdhury et al. [32] also implemented DenseNet201 and found it achieved an accuracy of 97.94% (when image augmentation was applied to the dataset) on the three-class problem, outperforming all the other models in the study in terms of the different performance indices. The architecture suggested by Gianchandani et al. [35] for the DenseNet201 model is shown in Figure 4 below.

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6	[1 × 1 conv, 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12	[1 × 1 conv, 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv, 3 × 3 conv] × 24	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 48	[1 × 1 conv, 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv, 3 × 3 conv] × 16	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 32	[1 × 1 conv, 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Figure 4: DenseNet201 Architecture [53]

In the DenseNet201 architecture, each layer is connected to every other layer. For every layer, the feature maps of all the preceding layers are used as inputs, and its own feature maps are used as input for each subsequent layer. This structure helps to alleviate the vanishing-gradient problem, strengthen feature propagation through the network, encourages feature reuse and greatly reduces the number of parameters, as stated in [20].

DarkNet, implemented by Chaddad et al. [23], was found to achieve an accuracy of 99.09% and AUC of 99.89% in classifying COVID-19 from non-COVID-19 (two-class problem), from a dataset that combined both X-Ray and CT scans. DarkNet was also found to achieve 97% overall accuracy in the three-class problem, including achieving 100% accuracy for the COVID-19 class.

Thus, it is suitable for the 3+-class study in this paper. The actual implemented version of DarkNet in this paper is DarkNet-19 [21]. The architecture of this model can be seen in Figure 5 below.

The DarkNet-19 architecture is similar to that of

Type	Filters	Size/Stride	Output
Convolutional	32	3 × 3	224 × 224
Maxpool		2 × 2/2	112 × 112
Convolutional	64	3 × 3	112 × 112
Maxpool		2 × 2/2	56 × 56
Convolutional	128	3 × 3	56 × 56
Convolutional	64	1 × 1	56 × 56
Convolutional	128	3 × 3	56 × 56
Maxpool		2 × 2/2	28 × 28
Convolutional	256	3 × 3	28 × 28
Convolutional	128	1 × 1	28 × 28
Convolutional	256	3 × 3	28 × 28
Maxpool		2 × 2/2	14 × 14
Convolutional	512	3 × 3	14 × 14
Convolutional	256	1 × 1	14 × 14
Convolutional	512	3 × 3	14 × 14
Convolutional	256	1 × 1	14 × 14
Convolutional	512	3 × 3	14 × 14
Maxpool		2 × 2/2	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	512	1 × 1	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	512	1 × 1	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	1000	1 × 1	7 × 7
Avgpool		Global	1000
Softmax			

Figure 5: DarkNet-19 Architecture [21]

VGG16. It uses 3x3 filters and doubles the number of channels at every pooling step, and global average pooling is used to make predictions. However, 1x1 filters are used to compress the feature representation between 3x3 convolutions.

The specific EfficientNet architecture implemented in this study was the EfficientNetB0 [22]. This is because this model was found to achieve 97.3% accuracy on a dataset containing 14124 X-Ray images of patients who had either COVID-19, pneumonia or were healthy, the three-class problem. The model also achieved 100% positive predicted value on COVID-19 detection. The architecture suggested by Tan et al. [22] for the EfficientNetB0 model is shown in Figure 6 below.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	28 × 28	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

Figure 6: EfficientNetB0 Architecture [22]

The EfficientNet architecture is a CNN architecture and scaling method that uniformly scales all dimensions (depth, width and resolution) using a compound coefficient. The EfficientNet scaling method uniformly scales network width, depth and resolution with a set of fixed scaling coefficients. To give an example, if we wanted to use 2^n times more computational resources, then we can simply increase the network depth by α^n, width by β^n,

and image size by γ^n, where α, β and γ are constant coefficients determined by a small grid search on the original small model. EfficientNet then uses ϕ, the compound coefficient, to uniformly scale the network width, depth and resolution in a principled way. The networks' compound scaling method is explained by the idea that if the input image is larger, then the network requires more layers to increase the receptive field, and more channels are needed to capture more fine-grained details in the larger image. The EfficientNetB0 architecture is based upon the mobile inverted bottleneck residual blocks (MBConv) of MobileNetV2 [54].

The Proposed Ensemble Model

The proposed ensemble model in this paper aims to build upon the success of the other state-of-the-art models in the 3-class problem. The architecture consists of a VGG16 model and an EfficientNetB0 model.

The VGG16 model was selected due to its performance in both accuracy and other metrics, as well as its good performance in the explainability metrics (mentioned in the next section). The EfficientNetB0 was chosen for the other model in the ensemble due to the fact that the EfficientNet architecture achieves state-of-the-art accuracy whilst being an order-of-magnitude smaller and faster than other contemporary models, as outlined in the paper by Tan et al. [22]. This allows for the model to be ran on hardware that doesn't require substantial memory or other high-end hardware, making it more suitable for mass-use in hospitals or other medical facilities where it can be utilised. Other papers in similar research problems have built ensemble models, such as the study by Gianchandani et al. [35] which built an ensemble method from VGG16 and DenseNet architectures for the 2-class COVID-19 prediction problem. However, due to the importance of run-time and computational resource use in this study, the EfficientNetB0 model was chosen as the second classifier in the ensemble. The EfficientNetB0 model was also found to achieve similar performance in both explainability and performance metrics to DenseNet201 in this study, and so the other benefits that EfficientNetB0 brings made it the better option for this ensemble method. One of the main reasons for this combination in the ensemble was due to the explainability results found for the VGG16 and EfficientNetB0 models, with each of the models making up for the downfalls of the other, and explained in the Results section below for the proposed ensemble model. Notably, the ensemble method proposed in this paper is able to be trained and used for inference upon a 16 gigabyte Graphic's Card, unlike an ensemble method using VGG16 and DenseNet201 models. These cards are much more affordable than a larger card required for these larger scale models.

C. Implementing Explainability

As stated previously, generating explanations for the predictions by the machine learning models is very important. The 4 state-of-the-art models implemented in this study do not generate these explanations, and so this functionality was built on top of them. Previously, two different methods for adding explainable predictions to machine learning models were discussed, LIME [25] and GradCAM [55]. In this study, GradCAM was selected to implement explainability into the predictions. This is because, with GradCAM, a heatmap can be produced indicating the areas of the image that were most important to the classifier in determining if the image showed signs of COVID-19 or not. Various studies, such as the study by Karim et al. [26] have shown that GradCAM can be implemented into VGG and DenseNet successfully for explainable COVID-19 predictions from chest X-Rays. Due to the similarity between the DarkNet architecture and the VGG16 architecture, this means that implementing GradCAM into DarkNet was also trivial. GradCAM has also been implemented into EfficientNet in many studies, such as the study by Bao Tram Duong [56], where brain tumour classification was investigated using an EfficientNet architecture and GradCAM for visualisations. The version of GradCAM used was just GradCAM, and not Guided GradCAM. This was due to the study by Jin et al. [46], which found Guided GradCAM did not capture all the disease tissue, and so lead to a lower performance.

GradCAM operates by looking at the final convolution layer of the network, and then examining the gradient information flowing into that layer. This means it is possible to visually validate where the network focuses its attention in the image, and so it can be verified that the model is looking at the correct patterns in the image. This is important as it allows the medical personnel to see why the model has come to the decision it has, and helps to build trust in the model's classifications if the medical personnel can validate the model is looking at the correct region in the image.

In order to implement GradCAM into the 5 models, the following target layers had to be extracted and their gradient information examined by GradCAM:

Table 1: GradCAM Target Layers

	VGG16	DenseNet201	DarkNet19	EfficientNetB0
Layer	model.features[-1]	model.features[-1]	model.conv5[-4]	model.features[-1]

In order to fully evaluate the model's heatmaps and their quality, metrics should be used, as well as qualitative analysis. Importantly, the diagnosis of respiratory illnesses from these medical scans is done by professionals, and so unless the results of these heatmaps was sent to medical professionals for quality assessment, qualitative analysis done in this paper would be sub-optimal, and lack scientific knowledge required for a valid analysis.

One such quantitative method for evaluating the heatmaps produced by the models is through attribution methods. These are explainability techniques that assign scores to each pixel in the image, based upon their importance to the classifier. Many different methods have been proposed in the literature, with the key idea being to remove pixels that are deemed to be the most important, and then reporting the drop in accuracy of the classifier on this resulting image compared to the original image. However, removing pixels can lead to the image being un-recognisable or even destroyed. One method for solving this issue is known as Remove and Debias (ROAD [57]). This operates by replacing a pixel with the weighted average of it's neighbours. This leads to a blurring effect on the image, instead of gaps or zeroed-out pixel values.

By specifying the percentile of pixels you wish to perturbate in the image, based upon their importance to the model, the percentage increase in the model's confidence in its classification can be calculated. This is done by measuring the confidence in the model's classification before the pixels are perturbated. Then, the pixels in the image are perturbated based upon their importance to the model in its classification. Then, the model tries to classify the resulting image after perturbation, and the change in the model's confidence in its prediction is calculated. In this paper, we used the Most Relevant First variation of the ROAD metric. This means that the pixels are perturbated in order from the most important pixels to the classifier being perturbated first. The higher the drop in the model's confidence, the better. This is because, by perturbating the most important pixels from the image, and then re-running the classification, the removal of these important pixels should mean that the model is now less confident in its classification.

By taking readings across different percentiles, we can determine how the classifier performs as more and more of the image is perturbated. However, a more robust metric can be created from taking the average confidence drop across all the percentile metrics. This metric would also be better for the comparison between the 5 models in this paper. The ROAD evaluation can also be done by removing the least important pixels determined by the models, and then re-running the classification on the remaining image. This technique is known as the Least Relevant First metric. This means that that the models should experience an increase in the confidence of their classification on this new image, as the pixels determined by the model to be of least importance for the classification are removed. By combining both Most Relevant First and Least Relevant First into a combined metric, known as ROAD Combined, we can create a better metric for comparison between the 5 models. The ROAD Combined formula is as follows:

$$Road\ Combined = \frac{(Least\ Relevant\ First\ Value - Most\ Relevant\ First\ Value)}{2}$$

This measures the increase in confidence for the model at a specified percentile for pixel importance.
By taking the average ROAD Combined score across different percentiles for each model, a final ROAD Combined drop in confidence value can be generated.

D. Model Evaluations

In order to evaluate the 5 different models and compare them against each other, different evaluation metrics need to be used. In this study, the evaluation metrics used were: accuracy, precision, recall, confusion matrices, sensitivity, specificity and time taken for model inference.

The output of all the 4 models in this study was a probability that a given image belongs to each of the three possible classes. This is then converted into a label via selecting the class with the highest probability. If the class with the highest probability is COVID-19, then this is the classification output of the model.

Accuracy was selected as it is important that the models predict an X-Ray or CT scan correctly. If a patient is not predicted to have COVID-19 by the model, but in actuality does have the COVID-19 disease, then this could mean that they may continue living their life as if they never contracted the disease. This could have repurcussions such as the patient spreading the disease to other people uknowingly, or not being as careful as they perhaps should be in avoiding contracting other respiratory illnesses in future due to potential lung scarring or other long-term effects of COVID-19.

Precision and recall are important metrics to use to compare models for this study. Precision measures the number of correct COVID-19 predictions out of all the datapoints predicted to be COVID-19 positive by the model. As such, it measures how trustworthy a positive COVID-19 prediction by a given model is. This trust is an important aspect to be investigated in this paper, as mentioned previously. Recall measures, out of all the COVID-19 positive scans in the dataset, how many the model predicted as being COVID-19 positive correctly. This is important as we want the models to correctly predict as many COVID-19 positive patients as possible, due to the potential long-term effects a positive infection can bring, as discussed previously.

The ROC curve is a metric used to compare the models in many related papers. The ROC curve provides the true positive rate as a function of false positive rate. The true positive rate is important in this study, as the models need to correctly predict all potential patients who are positive for COVID-19. The false positive rate is also an important metric as, if the model has a high false positive rate and predicts that a patient has the COVID-19 disease, but in actuality does not, then this could mean that a patient believes the symptoms they currently have are a result of COVID-19, when in reality they could be as a result of another illness or disease that needs to be diagnosed. However, these metrics and their meanings can be discovered through precision and recall, and hence the ROC curve is not used for the evaluation of models in this paper.

Due to the importance of the true positive and false positive rates, the confusion matrix was also used as an evaluation metric for this study. The confusion matrix also includes the true negative and true positive rates achieved by each model, and so also states how many predictions the models predicted correctly, an important metric.

The final two metrics used were sensitivity and specificity. Sensitivity measures the number of images correctly predicted as COVID-19 positive by the model out of the total number of actual COVID-19 positive images. This metric was thus also selected due to the same reasons as why it is important to measure accuracy. Specificity measures the number of correctly predicted non-COVID-19 images out of the total number of non-COVID-19 images. Due to the same reason as why the false positive rate is important, this is also an important metric, and so was selected for the study.

4 VALIDITY

The research question that this project addresses is: *What are the most effective machine learning models that can make explainable predictions on whether a CT or X-Ray scan shows evidence of a positive COVID-19 infection?*. In order for this research question to be acceptable, it must be

valid. As shown in the Related Work section, there are various other studies that have implemented similar work in aiming to detect COVID-19 infection from chest X-Ray or CT scans. Thus, as this study relates to existing similar validated studies, it fulfils concurrent validity. The mathematics behind deep learning and Convolutional Neural Networks has been proven many times over in literature, as well as their application to image analysis and classification tasks. As this study implements further features, such as explainable predictions and the ability to work on both X-Ray and CT images, on top of already proved theory, it also therefore has construct validity. By using the evaluation metrics stated in the previous section, as well as following the methodology clearly outlined in previous sections, this study does measure the most effective machine learning models for making explainable predictions on whether a CT scan or X-Ray scan shows evidence of a positive COVID-19 infection. This study also provides justification for all the choices in methodology, such as the models selected, the data used to create the dataset, as well as the evaluation metrics to compare the performances of the models. As a result, this study also has face validity, as it does appear to test what it aims to test. This study is also currently relevant, as it is using state-of-the-art machine learning models, and then building two additional features on top of them, namely, the ability to work on both X-Ray and CT scans, and to provide explainable predictions, something not included previously in other studies, as well as proposing a new ensemble model which outperforms these current models on the problem at hand. The benefits of this study are also outlined in the Introduction section of this paper. The reseach question is focused by narrowing in on one small area of COVID-19 prediction. Specifically, models that can predict the disease from both X-Rays and CT scans, as well as generating explainable predictions, which is also a cutting-edge area in the field.

The domain of this study is the prediction of COVID-19 from medical images, namely, X-Ray and CT scans. Many studies, such as [15] and [16] to name a few, have covered this domain, but many more studies have also been considered. This study also looks into building machine learning models that can operate on both X-Ray and CT scans. This domain was also investigated in literature, for example the study by Chaddad et al. [23]. Finally, the last domain this paper investigates is that of explainable predictions by machine learning classifiers. Two different potential solutions to this were found in the literature review, the LIME method outlined in [25] and GradCAM outlined in [55]. Thus, the related work section covers all the appropriate subject domains. The studies found in the Related Work section consists of papers that have been written after the year 2020. As COVID-19 is a relatively new disease, the investigation into automatic detection of the disease from medical images is also a new area of study. As a result, all papers referenced in this study are from the last few years, unless they are explaining the machine learning models or approaches that these papers have harnessed. This means that all papers referenced in this Related Work section are furthering research on a relatively new topic, and so these papers can be considered state-of-the-art. These models have achieved state-of-the-art performances in other classification tasks, such as VGG16 in the ImageNet Challenge. These models were also selected for this study due to their state-of-the-art performance in the task of detecting COVID-19 from medical images, with each model achieving an accuracy of over 97% in their respective studies.

As stated previously, the methodology to be implemented in this paper builds directly upon previous work in the field, thus providing construct and concurrent validity. The major sources for the implementation of the models, studies [19], [20], [21] and [22] have been cited 76423 times, 24025 times, 11505 times and 11667 times respectively. This solidifies the validity of the theory behind these models, as so many other papers have been reliant upon them. The extension ideas in this study, the ability for the models to operate on both X-Ray and CT images, and ensuring the models generate explainable predictions through GradCAM, are two ideas that have been used separately by other researchers previously, but not together. Using the same evaluation metrics and data to form the dataset as other studies also improves the validity of this study. This also allows for the findings of this study to be compared with those of these other studies. Additionally, the use of well-known evaluation metrics, such as accuracy, sensitivity and specificity to name a few, provides face validity.

5 RESULTS

Due to the fact that the datasets chosen for the combined dataset in this implementation were difficult images for classifiers, as well as the fact that image augmentation was applied in a way to replicate difficult medical images for classifiers, the overall performance of the models in this paper are slightly lower than the figures found in papers in similar research areas. This is also due to the fact that the accuracy of 3+ class classifiers can be lower than the accuracy of 3-class classifiers due to the fact that other respiratory illnesses may share common characteristics to COVID-19 infections, and so the classifier can mis-classify these images. The sections below outline the results found for each of the 5 models in this paper.

Each of the 5 models were trained for 100 epochs on a training set containing 70% of the total dataset size. A validation set containing 20% of the total dataset size was used to ensure overfitting did not occur. The test set was 10% of the total dataset size.

The ROAD metrics used to evaluate the heatmaps produced by each model are calculated by taking an average for each value at each percentile across 5 runs, and the ROAD Combined scores are calculated from taking an average across 5 different runs. This is to ensure a single run is not a bad representation of the model's performance.

A. VGG16

The following results were achieved by the VGG16 model on the entire test dataset.

Table 2: Classification Report for VGG16

Class	Precision	Recall	F1-score	Support
0	0.77	0.87	0.82	494
1	0.94	0.95	0.94	506
2	0.87	0.75	0.80	500

The VGG16 classifier was able to achieve an overall accuracy of 86% on the test set.
The following shows the confusion matrix for the VGG16 classifier on the test set:

Table 3: Confusion Matrix for VGG16

			Predicted	
		0 (Healthy)	1 (COVID-19)	2 (Other)
	0 (Healthy)	**432**	15	47
True	1 (COVID-19)	15	**480**	11
	2 (Other)	111	15	**374**

As can be seen above, the VGG16 classifier is able to predict all 3 classes with a high true positive rate. Notably, the classifier is able to achieve 94% precision and 95% recall on the COVID-19 class. A precision of 94% means that of all the datapoints that the classifier predicts as being COVID-19 positive, 94% of those were correct. This is important, as mentioned previously, as the classifiers have to be trustworthy. This means that if the classifier had a low precision, then if it was to predict a patient as having COVID-19, then it could mean there is a relatively high chance that the classifier is in fact incorrect. A recall of 95% means that of all the positive COVID-19 infections in the test set, the classifier was able to correctly predict 95% of them. This is important, as when the model is used in real-world inference, it needs to be able to predict patients who do in fact have COVID-19 correctly as much as possible.

The VGG16 classifier predicts a relatively large number of 'other' patient scans as the healthy class (111). In this paper, the primary aim is the correct classification of COVID-19 and so this is not critical. However, this means that, if this model is used to aid in accelerated diagnosis, it could mean the doctor may advise the patient that they do not have another respiratory disease that isn't COVID-19, such as pneumonia, when in fact they do. This could lead to patients believing that they are healthy when in actuality they are not, which could have serious consequences if a respiratory disease is left undiagnosed and untreated as a result.

Incorrect diagnosis is an important factor to consider when building trust in the model's outputs. As such, specificity is an important metric to consider. Specificity measures, out of all the scans that were not COVID-19 positive, how many did the classifier correctly predict. As stated above, an incorrect COVID-19 negative prediction can have serious adverse effects. For the VGG16 model, it achieved a specificity of 96%, which shows that it is capable at identifying images that were not COVID-19 positive.

The table below summarises all the evaluation metrics mentioned in this report:

Table 4: Evaluation Metrics Summary for VGG16

Evaluation Metric	
Accuracy	0.86
Precision (Class 1)	0.94
Recall (Class 1)	0.95
Sensitivity (Class 1)	0.95
Specificity (Class 1)	0.96
Time Taken on Test Set	24.2 seconds

Grad-CAM Heatmap

By using GradCAM, the following images show some example heatmaps that demonstrate the areas of the image that the model used to make its classification.

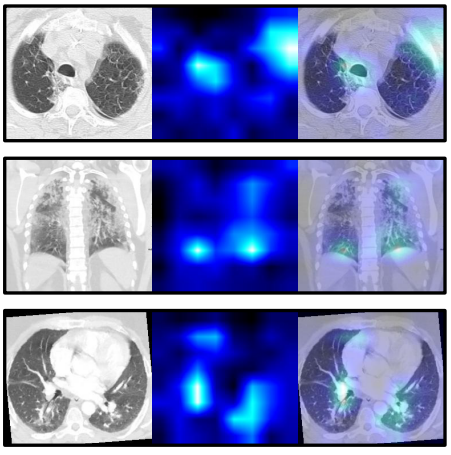

Figure 7: Example heatmap outputs from VGG16

The images above show the original image on the left, the heatmap generated by GradCAM in the middle, and the heatmap overlayed on top of the original image on the right.

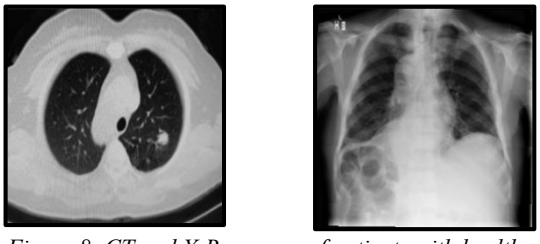

Figure 8: CT and X-Ray scans of patients with healthy lungs

The two images above in Figure 8 show the scans of a healthy individual's lungs. The image on the left is a CT scan and the image on the right is an X-Ray scan.

By comparing the heatmaps produced by GradCAM and these baseline images, it can be seen that the heatmaps highlight regions of the lungs that appear to be abnormal. For example, by enhancing the fourth heatmap image and the baseline CT scan of the healthy individual:

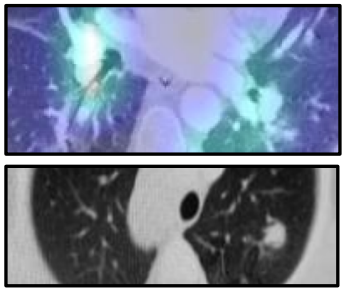

Figure 9: Highlighted regions by the VGG16 model for a positive COVID-19 infection

The top image in Figure 9 zooms in onto the highlighted region of the image. It can be seen in the patches highlighted the most vibrantly that there exist artefacts in the scans that do not appear in the same location on a healthy individual's lungs in the bottom image.

As mentioned previously, simply observing the heatmaps produced by each model is not sufficient to make an accurate and fair distinction between the classifiers' explainability performance. As such, for the remainder of this results section, we will focus upon the ROAD scores.

The graph below summarises the ROAD findings across 5 different percentiles for the VGG16 model:

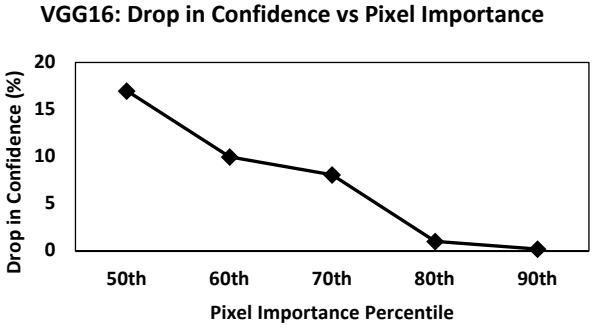

Figure 10: ROAD Most Relevant First scores across various percentiles of pixel importances for VGG16

As shown, as we perturbate pixels in higher importance percentiles, the classifier's confidence in its decision falls. This is because, when you perturbate only the top 10% of pixels that are important to the model in its classification, less pixels are being removed from the image. This means the image is closer to the original image than if you remove the top 50% of important pixels to the model. As such, the model is still accessing most of the data in the original image, and so is likely to still make the same classification prediction.

By using the more robust ROAD Combined metric specified previously, we can produce a metric that can be better compared to the other models.

The Average ROAD Combined across 5 percentiles for VGG16: 4.35

B. DenseNet201

The following results were achieved by the DenseNet201 model on the entire test dataset.

Table 5: Classification Report for DenseNet201

Class	Precision	Recall	F1-score	Support
0	0.79	0.86	0.82	493
1	0.95	0.97	0.96	497
2	0.84	0.75	0.79	510

The DenseNet201 model was able to achieve an overall accuracy of 86% on the test set.
The following shows the confusion matrix for the DenseNet201 classifier on the test set:

Table 6: Confusion Matrix for DenseNet201

			Predicted	
		0 (Healthy)	1 (COVID-19)	2 (Other)
	0 (Healthy)	**422**	6	65
True	1 (COVID-19)	5	**482**	10
	2 (Other)	107	18	**385**

The confusion matrix above shows that the DenseNet201 architecture was also able to have a high number of true positives across all 3 classes. However, the classifier incorrectly predicted 65 healthy scans as 'other' scans. This is an approximately 50% increase on the value found for the VGG16 model. This means the model is predicting the patient to have another respiratory disease, when in fact they show no evidence of having such respiratory disease. This could lead to the patient going for a check-up or follow-up scans for a respiratory disease due to the fact that the model told them their lung scan wasn't healthy, when in fact they do not need such an appointment. This could reduce trust in the model's predictions.

However, the DenseNet201 architecture predicted 506 patients as being COVID-19 positive and was incorrect on only 24 of these predictions. This means that the model had a precision of 95%. This is important as it builds trust in the model's predictions for the patients. The DenseNet201 classifier also achieved a specificity of 97%, which means the classifier's output on non-COVID-19 positive scans can be trusted as well, particularly as this score is also higher than that achieved by the VGG16 classifier. The classifier predicted 482 of the 497 true COVID-19 patients in the test set, meaning it achieved a recall of 97%. This is an improvement in both the precision and recall results found by the VGG16 model, and for the reasons mentioned previously, is an important factor to consider in the comparison between these models. The table below summarises all the evaluation metrics mentioned in this report:

Table 7: Evaluation Metrics Summary for DenseNet201

Evaluation Metric	
Accuracy	0.86
Precision (Class 1)	0.95
Recall (Class 1)	0.97
Sensitivity (Class 1)	0.97
Specificity (Class 1)	0.97
Time Taken on Test Set	30.4 seconds

Grad-CAM Heatmap

As can be seen in the graph below, DenseNet201's drop in confidence as we increase the percentile falls from 8.37% for the 60th percentile to 3.85% for the 90th percentile.

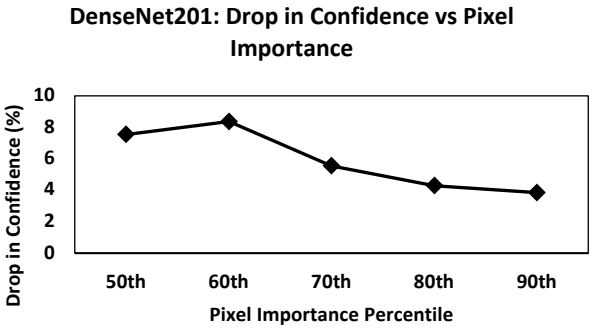

Figure 11: ROAD Most Relevant First scores across various percentiles of pixel importances for DenseNet201

This means that the DenseNet201 witnessed smaller drops in confidence as a result of perturbating the pixels it deemed as the most important in the image. This would suggest that the predictions made by this model are less explainable than those made by the VGG16 model. However, the graph above also shows that when just the top 90th percentile of pixels are perturbated in the image, the confidence drop of the classifier is still 3.85%. This is much higher than the value found for VGG16, and could suggest that the DenseNet201 classifier is better at detecting these most important pixels for classification, but less capable at predicting the less important pixels for classification.

The average ROAD Combined across 5 percentiles for DenseNet201: 3.70

This shows that the DenseNet201 model can produce heatmaps for its explanations that lead to a slightly lower increase in confidence using the ROAD Combined metric than the VGG16 model. Additionally, this backs up the hypothesis that the DenseNet201 model is less capable than VGG16 at predicting the less most important pixels for classification. This is because the ROAD Combined metric takes into account both Most Relevant First and Least Relevant First pixels in the image to calculate its scores. As the graph shows the DenseNet201 model appears to be better, or at least very similar, at detecting the most important pixels in the image, a lower ROAD Combined score than VGG16 suggests it is in fact worse at detecting the least relevant pixels in the image.

C. DarkNet19

The following results were achieved by the DarkNet19 model on the entire test dataset.

Table 8: Classification Report for DarkNet19

Class	Precision	Recall	F1-score	Support
0	0.84	0.77	0.80	500
1	0.96	0.92	0.94	486
2	0.77	0.87	0.82	514

The DarkNet19 model achieved an overall accuracy of 85% on the test set.
The confusion matrix for the DarkNet19 model can be seen below:

Table 9: Confusion Matrix for DarkNet19

			Predicted	
		0 (Healthy)	1 (COVID-19)	2 (Other)
	0 (Healthy)	**384**	10	106
True	1 (COVID-19)	13	**449**	24
	2 (Other)	62	7	**445**

The classification report above shows that the DarkNet19's precision of 96% was better than both VGG16 and DenseNet201. This can be seen in the confusion matrix by the fact that out of all the scans predicted to be COVID-19 positive in the test dataset, 10 of the scans were in fact healthy and 7 were other respiratory diseases. However, the recall for the COVID-19 class for the DarkNet19 model was only 92%, worse than the results achieved by the VGG16 model and DenseNet201 model. This study places equal emphasis on both precision and recall, due to previously mentioned reasons, and so this drop in recall for this model is important. However, the DarkNet19 model achieved a specificity of 98%, higher than both VGG16 and DenseNet201 models, meaning it is more capable at identifying correctly non-COVID-19 scans.
The table below summarises all the evaluation metrics mentioned in this report:

Table 10: Evaluation Metrics Summary for DarkNet19

Evaluation Metric	
Accuracy	0.85
Precision (Class 1)	0.96
Recall (Class 1)	0.92
Sensitivity (Class 1)	0.92
Specificity (Class 1)	0.98
Time Taken on Test Set	20.8 seconds

Grad-CAM Heatmap

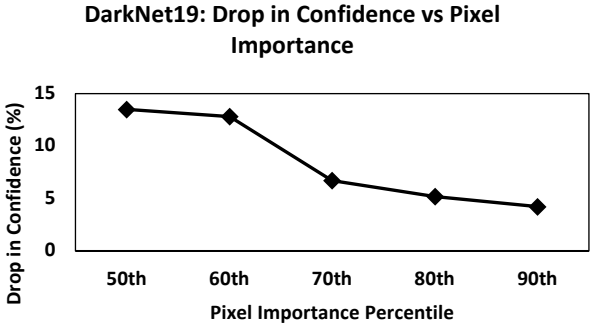

Figure 12: ROAD Most Relevant First scores across various percentiles of pixel importances for DarkNet19

As can be seen in the graph above, the DarkNet19 model was able to achieve between 13.5% and 4.23% drop in confidence between perturbating the 50th percentile of important pixels up to the 90th percentile. This is a similar drop in confidence to the VGG16 model, but performs better for the 80th and 90th percentiles. The 70th, 80th and 90th percentiles achieve similar performance to DenseNet201 but better for the 50th and 60th. However, DarkNet19 achieves slightly worse performance for the 50th and 60th percentiles compared to VGG16.

The average ROAD Combined across the 5 percentiles for DarkNet19: 4.89

This means that the DarkNet19 model is able to achieve an increase in confidence in its classification when pixels are perturbated from the image based upon their importance judged by the classifier by more than the VGG16 and DenseNet201 classifiers. This suggests the DarkNet19 model can produce more explainable predictions than these other 2 models.

D. EfficientNetB0

The following results were achieved by the EfficientNetB0 model on the entire test dataset.

Table 11: Classification Report for EfficientNetB0

Class	Precision	Recall	F1-score	Support
0	0.81	0.76	0.78	503
1	0.96	0.93	0.94	490
2	0.76	0.83	0.79	507

The EfficientNetB0 model achieved an overall accuracy of 84% on the test set.
The confusion matrix for the EfficientNetB0 model can be seen below.

Table 12: Confusion Matrix for EfficientNetB0

			Predicted	
		0 (Healthy)	1 (COVID-19)	2 (Other)
	0 (Healthy)	**380**	9	114
True	1 (COVID-19)	14	**454**	22
	2 (Other)	77	10	**420**

For the COVID-19 class, the EfficientNetB0 model had a precision of 96%and recall of 93%. This is because out of all the samples that were predicted to be COVID-19 positive, the classifier correctly predicted 454 out of 473 datapoints, giving a precision of 96%. Out of all the datapoints that were in fact COVID-19 positive, 454 out of the 490 datapoints were predicted correctly by the classifier, and so the recall was 93%. The EfficientNetB0 model matches the highest precision found by the other 3 models, but does not have the highest recall. The EfficientNetB0 also matches the highest specificity result found by the models so far, with 98%, meaning it not only has trustworthy COVID-19 predictions indicated by its high precision value, but also high reliability in its non-COVID-19 scan classifications.
The table below summarises all the evaluation metrics mentioned in this report:

Table 13: Evaluation Metrics Summary for EfficientNetB0

Evaluation Metric	
Accuracy	0.84
Precision (Class 1)	0.96
Recall (Class 1)	0.93
Sensitivity (Class 1)	0.93
Specificity (Class 1)	0.98
Time Taken on Test Set	20.8 seconds

Grad-CAM Heatmap

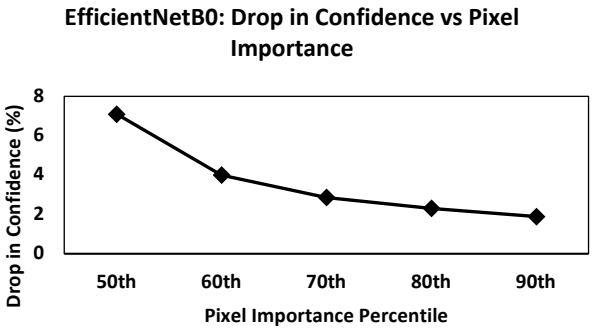

Figure 13: ROAD Most Relevant First scores across various percentiles of pixel importances for EfficientNetB0

As can be seen in the graph above, the EfficientNetB0 model achieves a maximum of 7.23% drop in confidence as a result of perturbating the top 50% most important pixels in the image. This is lower than the VGG16 and DenseNet201 models, but higher than the DarkNet19 model.

The average ROAD Combined across 5 percentiles for EfficientNetB0: 4.21

This ROAD Combined score and the graph showing confidence drop vs pixel importance are close to the performance achieved by the VGG16, thus suggesting the heatmaps produced by this model are of high quality and can be used to generate explainable classifications.

E. Proposed Ensemble Model

The following results were achieved by the proposed ensemble model in this paper:

Table 14: Classification Report for Proposed Ensemble Method

Class	Precision	Recall	F1-score	Support
0	0.80	0.85	0.82	505
1	0.99	0.97	0.98	496
2	0.83	0.80	0.81	499

The proposed hierarchical architecture is able to achieve 87% accuracy on the test set. This is a higher accuracy performance than the other 4 state-of-the-art models on the test set. The confusion matrix can be seen below:

Table 15: Confusion Matrix for Proposed Ensemble Method

			Predicted	
		0 (Healthy)	1 (COVID-19)	2 (Other)
	0 (Healthy)	**427**	5	73
True	1 (COVID-19)	7	**482**	7
	2 (Other)	100	2	**397**

This classification report shows that the proposed hierarchical model has a recall of 97%, and a precision of 99% for the COVID-19 positive class. As stated previously, precision and recall are important metrics to consider in this paper's investigation. Notably, out of all the predicted COVID-19 positive medical scans in the test set, the model predicted only 7 incorrectly, out of a total of 489 images, giving the precision of 99%. Additionally, out of all the images that were actually COVID-19 positive, the model only incorrectly predicted 14 out of 496. However, this model predicts a relatively high number of false positives for the healthy and 'other' classes. In particular, the model predicts 'other' scans as belonging to the healthy class for 100 samples out of the 534 total samples predicted as belonging to the healthy class, and 73 healthy scans as belonging to the 'other' class out of a total of 477 total predicted 'other' scans in the test set. However, despite these numbers of incorrect predictions appear in-line with the other state-of-the-art models, this proposed ensemble model achieves a specificity of 99%, higher than all 4 of the state-of-the-art models. This means that its classifications on non-COVID-19 scans is more accurate than these other 4 models. A summary of this model's stats can be found below:

Table 16: Evaluation Metrics Summary for Proposed Ensemble Method

Evaluation Metric	
Accuracy	0.87
Precision (Class 1)	0.99
Recall (Class 1)	0.97
Sensitivity (Class 1)	0.97
Specificity (Class 1)	0.99
Time Taken on Test Set	49.6 seconds

As can be seen above, this model takes approximately 2 times the time to run on the test set as the other models in this study. However, 50 seconds is still fast enough to be used in the real-world and not have doctors waiting for the results of the model inference for too long in order to aid with the diagnosis. This is because the test set contains 1500 samples and so 50 seconds to run on all these samples means it takes a minor amount of time to run on any individual sample.

Grad-CAM Heatmap

In order to create heatmaps with the ensemble model, the target layers for both the VGG16 and EfficientNetB0 layers are used. The result of these target layers for each of the models in the ensemble is then averaged to produce the final output for the whole ensemble model. This brings the explainability benefits from both models together to produce a better heatmap. Notably, the EfficientNetB0 achieves higher ROAD Most Relevant First scores for the 80th and 90th percentile of pixel importances. This makes up for the downfall of the VGG16 model, which achieves very low drop in confidences at these percentiles. Additionally, the VGG16 model achieves higher drops in confidence for the 50th and 60th percentiles, where the EfficientNetB0 model performs less well. These findings can be seen in Figures 10 and 13. Thus, the proposed ensemble model brings the benefits of both classifiers together in order to bring substantial performance benefits in explainability compared to the 4 other state-of-the-art models, as can be seen in Figure 14 below.

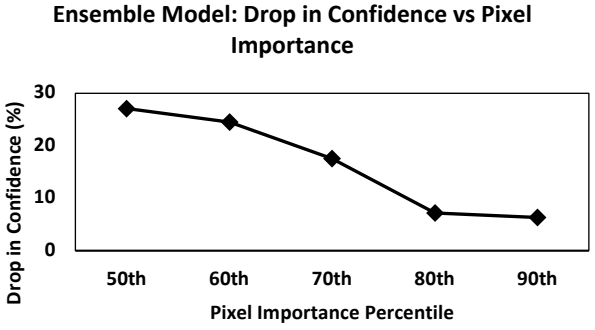

Figure 14: ROAD Most Relevant First scores across various percentiles of pixel importances for Proposed Ensemble Method

As can be seen in the graph above, the proposed ensemble model's confidence drops from 27.1% at the 50th percentile of important pixels to 6.32% at the 90th percentile pixels. This is a significant increase from the other 4 state-of-the-art models tested in this paper, which found the highest confidence drop at the 90th percentile to be 4.23% and the highest confidence drop at 50th percentile to be 17% for the DarkNet19 model and VGG16 model respectively.

In order to properly compare against the other 4 models across all 5 percentiles, the ROAD Combined score for the proposed model was calculated.

Average ROAD Combined across 5 percentiles: 9.64

This ROAD Combined score is also significantly above the other 4 state-of-the-art models, with the highest ROAD Combined score before this proposed model being 4.89.

6 EVALUATION

To evaluate the models in this paper, and to answer the research question, the models have to compared both on performance metrics and explainability metrics.

In this study, the models' performance was measured using the following metrics: accuracy, precision, recall, time taken for inference on test set, sensitivity and specificity. The table below summarises the results for each of the models on these metrics:

Table 17: Evaluation Metrics Comparison with all 5 models

Evaluation Metric	VGG16	DenseNet201	DarkNet19	EfficientNetB0	**Our Ensemble Model**
Accuracy	0.86	0.86	0.85	0.84	**0.87**
Precision (Class 1)	0.94	0.95	0.96	0.96	**0.99**
Recall (Class 1)	0.95	**0.97**	0.92	0.93	**0.97**
Sensitivity (Class 1)	0.95	**0.97**	0.92	0.93	**0.97**
Specificity (Class 1)	0.96	0.97	0.98	0.98	**0.99**
Time Taken on Test Set (seconds)	24.2	30.4	**20.8**	**20.8**	49.6

Highlighted in bold in the table above are the models that achieved the highest result for each of the evaluation metrics. As can be clearly seen, the proposed ensemble method in this paper achieves the highest performance in all the evaluation metrics apart from the time taken for the model to run its inference on the test set. The test set contains 1500 samples, and so if these models were ran on a singular medical image in reality, then the time taken per scan is still insignificant, even if our proposed model takes twice the length of time on the 1500 sample test set. As such, the drawback of our proposed model on this metric is greatly outweighed by its performances across the other metrics. The capability of the proposed ensemble method at predicting both COVID-19 positive scans more accurately (shown by its precision in the COVID-19 positive class) and its ability to accurately predict non-COVID-19 positive scans (shown by its high specificity) shows that the model's classification outputs can also be trusted more than those produced by the other 4 state-of-the-art models.

As our proposed ensemble method out-performs all the other state-of-the-art models across all metrics apart from the time taken for inference (which has been explained to be less important), this model is therefore the most suitable model for hybrid CT/X-Ray can COVID-19 prediction in terms of classification performance.

As stated previously, explainability is a very important aspect to be considered in the medical domain. As such, in order to answer the research question and determine the overall best model for this use case, the models have to be evaluated for their ability to provide explainable predictions. In this paper, we used ROAD Most Relevenant First across different percentiles of pixel importances, and a custom ROAD Combined metric, which provided a more robust method for comparison of the model's explainability across the 5 models in this paper. The graph below shows the 5 different models and their drop in confidence against the 5 different percentiles of pixel importances.

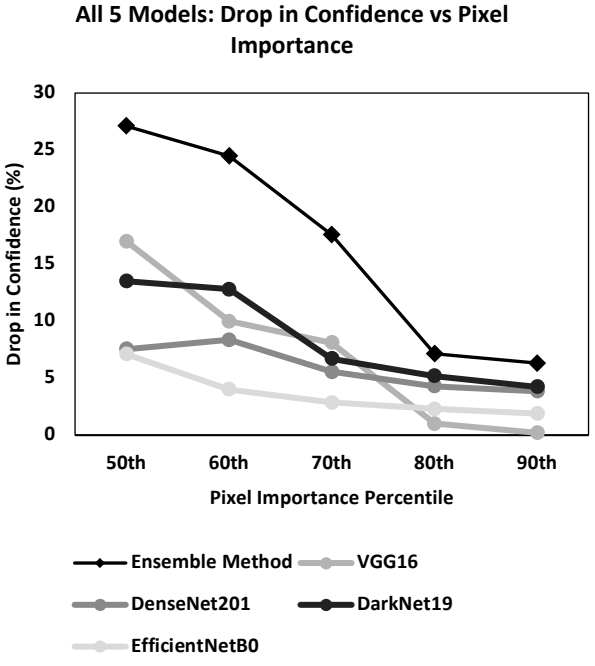

Figure 15: ROAD Most Relevant First scores for all 5 models across various percentiles of pixel importances

As can be seen in Figure 15 above, our proposed ensemble method provides a greater percentage drop in confidence in its classifications for each of the 5 different percentile of pixel importances, as shown by the dark blue graph above. This means that as the most important pixels deemed by the classifier itself are perturbated from the image, and the classification ran again, the classifier suffers the greatest drop in confidence in its classification for our proposed ensemble method. This means that the most important pixels identified by the ensemble model are in fact the most influential pixels in the image for a correct COVID-19 classification. This implies that our proposed model is in fact able to produce the most explainable predictions out of the 5 models in this paper, as the regions identified in its heatmap produced by GradCAM are the most important regions for correct classifications.

To further compare the models' explainability capabilities, the ROAD Combined metric was created. The table below summarises the findings for each of the 5 models.

Table 18: ROAD Combined scores for all 5 models

	VGG16	Dense-Net201	Dark-Net19	Efficient-NetB0	Proposed Ensemble Model
ROAD Combined Score	4.35	3.70	4.89	4.21	**9.64**

The ROAD Combined metric takes the average of the Least Relevant First and Most Relevant First ROAD scores at a particular percentile. As such, it takes into account both the most important and least important pixels identified by the model in the image, instead of just the most important pixels in the ROAD Most Relevant First scores shown in Figure 15 above. A higher score for the ROAD Combined means that a model is better able to identify the least important and most important pixels in the image.

As can be seen in the table, the 4 state-of-the-art models achieved a score of approximately 4 for ROAD Combined. Whereas, the proposed ensemble model in this paper achives a score of 9.64. This is approximately double the score of any of the 4 state-of-the-art models and as such shows the proposed model in this paper performs best at explainability by this metric as well.

As the proposed ensemble model in this paper outperforms the 4 state-of-the-art models in both performance metrics as well as explainability metrics, it is therefore the most suitable model to make explainable predictions on whether a CT or X-Ray scan shows evidence of a positive COVID-19 infection, answering the research question.

7 CONCLUSION

In this paper, 4 state-of-the-art machine learning models are selected based upon their performances in similar research areas in order to fulfil the task of COVID-19 infection detection from a hybrid dataset containing both CT and X-Ray scans. In particular, the models are trying to, given a medical image that could be either a CT scan or an X-Ray scan of a patient's lungs, identify if the image belongs to one of 3 classes: healthy, COVID-19 positive, or 'other' respiratory diseases. In this paper, the 'other' class contains an array of various different respiratory diseases, such as pneumonia. As this 'other' class contains multiple different respiratory diseases that can share different characteristics in the medical scans, this research problem is known as a 3+ class problem, as there are effectively at least 3 different classes that a medical image could belong to.

Then, as most of these models in the literature do not provide explainable predictions, this capability is built upon these models, through using GradCAM [55]. This allows each of the models to provide a heatmap indicating the areas of the image that they focussed upon in order to make their classification. This can then be qualitatively assessed through certified medical personnel checking if these areas of the image actually represent regions indicative of a COVID-19 infection, as well as quantitively through the use of state-of-the-art metrics such as Remove and Debias (ROAD) [57]. In this paper, we evaluated and compared the models using ROAD Most Relevant First and a custom ROAD Combined metric. This operates by taking the most important pixels identified by a given classifier, and blurring them in the image (known as perturbating the pixels). The classifier is then run upon this new image with the perturbed pixels. Then, the model's confidence in this new classification is compared to its confidence on the original image, and the drop in confidence calculated. The better the classifier's ability to provide explainable predictions, the better its ability to identify these most important pixels in the image, and so after these pixels are blurred, its confidence at classifying this modified image should fall significantly. As such, the more explainable the classifier, the greater the drop in confidence. For the ROAD Combined metric, it takes into account the least important pixels identified in the image as well as the most important pixels identified in the image for a given classifier, with the higher the ROAD Combined score, the better the classifier at providing explainable predictions.

Then, a new ensemble method is proposed which builds upon the advances in these models. In particular, by taking into account the outputs of 2 of the state-of-the-art classifiers (VGG16 and EfficientNetB0), it can create a more accurate classification. Additionally, by averaging the heatmaps produced by both these models, it takes the advantages of both VGG16's explainability and EfficientNetB0's explainability, whereby VGG16 is better at determing the top 50% of important pixels, but less capable at identifying the top 20% of important pixels, and EfficientNetB0 is more accurate at identifying these top 20% of important pixels.

This means the proposed ensemble method is capable of outperforming these state-of-the-art models in both performance metrics such as accuracy, precision, recall, specificity and sensitivity as well as significant improvements versus state-of-the-art in both explainability metrics (ROAD Most Relevant First and ROAD Combined) used in this paper.

There are some limitations of the methods followed in this paper, that could be built upon in future work. The dataset created in this paper relies heavily upon image augmentation to create a dataset of size 15000 samples. However, if more original data could be used to create a dataset of equal size or even one of larger size by also utilising image augmentation, this could create models that achieve better performance on out-of-sample datapoints. Additionally, the use of Remove and Debias [55] to evaluate heatmaps produced by the models is only one method of evaluating the explainability of models, and the comparison of the models' capabilities could be further investigated through the use of multiple metrics being used concurrently. As found in this paper, all the models achieved very high performance on the COVID-19 positive class, but reduced performance on the healthy and 'other' classes. In particular, the models tended to predict healthy scans as 'other' and 'other' scans as healthy. Whilst this could be explained by the fact that the 3+ class problem can give rise to multiple different respiratory diseases which may have subtle indicators in medical images as to their diagnosis, this issue could be improved upon by potentially changing the weightings in the loss functions of the models or by providing more data, as mentioned previously.

REFERENCES

[1] World Health Organization, 2020. Naming the coronavirus disease (COVID-19) and the virus that causes it. *Brazilian Journal Of Implantology And Health Sciences*, *2*(3).

[2] AdhanomT, W.H.O., 2020. Director-General's Opening Remarks at the Media Briefing on COVID [WHO web site]. 2020 Available at: https://www. who. int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. *Accessed May*, *31*.

[3] Lippi, G., Nocini, R., Mattiuzzi, C. and Henry, B. (2021) Is body temperature mass screening a reliable and safe option for preventing COVID-19 spread?. Diagnosis, Vol. (Issue), pp. 000010151520210091. https://doi.org/10.1515/dx-2021-0091

[4] Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F. and Liu, J., 2020. Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing. *Radiology*, *296*(2), pp.E41-E45.

[5] Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P. and Ji, W., 2020. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology*, *296*(2), pp.E115-E117.

[6] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z. and Xia, L., 2020. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*, *296*(2), pp.E32-E40.

[7] American College of Radiology, 2020. ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection.

[8] Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., Cui, J., Xu, W., Yang, Y., Fayad, Z.A. and Jacobi, A., 2020. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology*, *295*(1), pp.202-207.

[9] Li, X., Zeng, W., Li, X., Chen, H., Shi, L., Li, X., Xiang, H., Cao, Y., Chen, H., Liu, C. and Wang, J., 2020. CT imaging changes of corona virus disease 2019 (COVID-19): a multi-center study in Southwest China. *Journal of translational medicine*, *18*(1), pp.1-8.

[10] Bai, H.X., Hsieh, B., Xiong, Z., Halsey, K., Choi, J.W., Tran, T.M.L., Pan, I., Shi, L.B., Wang, D.C., Mei, J. and Jiang, X.L., 2020. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology*, *296*(2), pp.E46-E54.

[11] Salehi, S., Abedi, A., Balakrishnan, S. and Gholamrezanezhad, A., 2020. Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. *Ajr Am J Roentgenol*, *215*(1), pp.87-93.

[12] Salehi, S., Abedi, A., Balakrishnan, S. and Gholamrezanezhad, A., 2020. Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. *Ajr Am J Roentgenol*, *215*(1), pp.87-93.

[13] Kligerman, S., Raptis, C., Larsen, B., Henry, T.S., Caporale, A., Tazelaar, H., Schiebler, M.L., Wehrli, F.W., Klein, J.S. and Kanne, J., 2020. Radiologic, pathologic, clinical, and physiologic findings of electronic cigarette or vaping product use–associated lung injury (EVALI): evolving knowledge and remaining questions. *Radiology*, *294*(3), pp.491-505.

[14] Abdollahi, B., El-Baz, A. and Frieboes, H.B., 2019. Overview of deep learning algorithms applied to medical images. *Big Data in Multimodal Medical Imaging*, pp.225-237.

[15] Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q. and Cao, K., 2020. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology*, *296*(2), pp.E65-E71.

[16] Luz, E., Silva, P., Silva, R., Silva, L., Guimarães, J., Miozzo, G., Moreira, G. and Menotti, D., 2021. Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images. *Research on Biomedical Engineering*, pp.1-14.

[17] Sun, Q., Lin, X., Zhao, Y., Li, L., Yan, K., Liang, D., Sun, D. and Li, Z.C., 2020. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. *Frontiers in oncology*, *10*, p.53.

[18] https://www.bbc.co.uk/news/uk-60467183

[19] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[20] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[21] Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271).

[22] Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR

[23] Chaddad, A., Hassan, L. and Desrosiers, C., 2021. Deep CNN models for predicting COVID-19 in CT and x-ray images. *Journal of medical imaging*, *8*(S1), p.014502.

[24] Ethics of AI in Radiology: European and North American Multisociety Statement. American College of Radiology, 2020.

[25] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[26] Karim, M., Döhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M. and Beyan, O., 2020. Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images. *arXiv preprint arXiv:2004.04582*.

[27] Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O. and Acharya, U.R., 2020. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in biology and medicine*, *121*, p.103792.

[28] Chouhan, V., Singh, S.K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R. and De Albuquerque, V.H.C., 2020. A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, *10*(2), p.559

[29] Gu, X., Pan, L., Liang, H. and Yang, R., 2018, March. Classification of bacterial and viral childhood pneumonia using deep learning in chest radiography. In *Proceedings of the 3rd international conference on multimedia and image processing* (pp. 88-93).
[30] Minaee, S., Kafieh, R., Sonka, M., Yazdani, S. and Soufi, G.J., 2020. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical image analysis*, *65*, p.101794.
[31] https://github.com/ieee8023/covid-chestxray-dataset
[32] M. E. H. Chowdhury *et al*., "Can AI Help in Screening Viral and COVID-19 Pneumonia?," in *IEEE Access*, vol. 8, pp. 132665-132676, 2020, doi: 10.1109/ACCESS.2020.3010287.
[33] Perez, L. and Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.
[34] Das, N.N., Kumar, N., Kaur, M., Kumar, V. and Singh, D., 2020. Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays. *Irbm*.
[35] Gianchandani, N., Jaiswal, A., Singh, D., Kumar, V. and Kaur, M., 2020. Rapid COVID-19 diagnosis using ensemble deep transfer learning models from chest radiographic images. *Journal of ambient intelligence and humanized computing*, pp.1-13.
[36] Alakus, T.B. and Turkoglu, I., 2020. Comparison of deep learning approaches to predict COVID-19 infection. *Chaos, Solitons & Fractals*, *140*, p.110120.
[37] Chaddad, A., Toews, M., Desrosiers, C. and Niazi, T., 2019. Deep radiomic analysis based on modeling information flow in convolutional neural networks. *IEEE Access*, *7*, pp.97242-97252.
[38] https://towardsdatascience.com/investigation-of-explainable-predictions-of-covid-19-infection-from-chest-x-rays-with-machine-cb370f46af1d
[39] Aditya Chattopadhay and Anirban Sarkar. 2018. Grad-CAM++: Generalized gradient-based visual explanations for convolutional networks. In Applications of Computer Vision(WACV). IEEE, 839–847.
[40] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. 2019. Explaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation. arXiv:1908.04351 (2019).
[41] M. R. Karim, M. Cochez, O. Beyan, S. Decker, and C. Lange. 2019. OncoNetExplainer: Explainable Predictions of Cancer Types Based on Gene Expression Data. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). 415–422.
[42] Mei, X., Lee, H.C., Diao, K.Y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M. and Bernheim, A., 2020. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature medicine*, *26*(8), pp.1224-1228
[43] Bai, H.X., Wang, R., Xiong, Z., Hsieh, B., Chang, K., Halsey, K., Tran, T.M.L., Choi, J.W., Wang, D.C., Shi, L.B. and Mei, J., 2020. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology*, *296*(3), pp.E156-E165.
[44] Wehbe, R.M., Sheng, J., Dutta, S., Chai, S., Dravid, A., Barutcu, S., Wu, Y., Cantrell, D.R., Xiao, N., Allen, B.D. and MacNealy, G.A., 2021. DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical data set. *Radiology*, *299*(1), pp.E167-E176.
[45] Murphy, K., Smits, H., Knoops, A.J., Korst, M.B., Samson, T., Scholten, E.T., Schalekamp, S., Schaefer-Prokop, C.M., Philipsen, R.H., Meijers, A. and Melendez, J., 2020. COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. *Radiology*, *296*(3), pp.E166-E172.
[46] Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., Deng, L., Zheng, C., Zhou, J., Shi, H. and Feng, J., 2020. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nature communications*, *11*(1), pp.1-14.
[47] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2017, http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf
[48] Angelov, P. and Almeida Soares, E., 2020. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. *MedRxiv*.
[49] Zhao, J., Zhang, Y., He, X. and Xie, P., 2020. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, *490*.
[50] Aswathy, A.L., Hareendran, A. and SS, V.C., 2021. COVID-19 diagnosis and severity detection from CT-images using transfer learning and back propagation neural network. *Journal of Infection and Public Health*, *14*(10), pp.1435-1445
[51] Tompe, A. and Sargar, K., 2021. X-Ray Image Quality Assurance. In *StatPearls [Internet]*. StatPearls Publishing.
[52] Nash, W., Drummond, T. and Birbilis, N., 2018. A review of deep learning in the study of materials degradation. *npj Materials Degradation*, *2*(1), p.37.
[53] https://pytorch.org/hub/pytorch_vision_densenet/
[54] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).
[55] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
[56] https://medium.com/mlearning-ai/explainable-ai-brain-tumor-classification-with-efficientnet-and-gradient-weighted-class-activation-24c57ae6175d
[57] Rong, Y., Leemann, T., Borisov, V., Kasneci, G. and Kasneci, E., 2022. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*.