.

INVESTIGATION ON THE EFFECTS OF RADIOGRAPHIC IMAGE QUALITY ATTRIBUTES ON THE PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORKS (CNNS) IN DETECTING COVID-19

REYMOND R. MESUGA[1], CLOYD RAYMOND P. PERNES[1], LUTHER A. VILLACRUZ[1,2], AND MARK ANTHONY C. BURGONIO[1,3]

[1]Department of Physical Sciences, College of Science, Polytechnic University of the Philippines, Sta. Mesa, Manila 1016, Philippines
[2]Department of Physics, School of Science and Engineering, Ateneo de Manila, Loyola Heights, Quezon City 1108, Philippines
[3]Cebu Cancer Institute, Perpetual Succour Hospital of Cebu, Inc., Cebu City 6000, Philippines

Abstract: Radiographic image quality is one of the factors that impacts professionals' decisions when diagnosing lung diseases using X-ray images. Hence, poor radiographic image quality could result in a misleading diagnosis affecting the person being investigated. This is true in human vision, as well as the computer vision. This study investigated the effects of different radiographic image quality attributes (i.e., contrast, Gaussian blur, Gaussian noise, and salt-and-pepper noise) on the performance of various Convolutional Neural Networks (CNNs) models. We use COVID-19 x-ray data as an initiative to the pandemic, apply different radiographic image quality attributes, and test the performance of CNN models in the effects of the attributes in the classification task. The results showed the following: (i) increasing levels of experimented noises (i.e., Gaussian and salt-and-pepper noise) rapidly decreases the performance of the models with no sign of resiliency; (ii) decreasing contrast appears to be beneficial at some particular level (e.g., contrast factor = 3); and (iii) increasing Gaussian blur decreases the performance of models but less rapidly than that of noises. As a conclusion, increasing noise like Gaussian and salt-and-pepper noise can be considered as a hindrance to the performance of CNNs while decreasing contrast and increasing Gaussian blur seemed to be beneficial especially if applied for data augmentation or enhancement techniques as the performance of the CNNs were observed to be more resilient against these two attributes than that of noises.

Keywords: deep learning (DL), convolutional neural network (CNN), COVID-19, radiographic image quality

1. INTRODUCTION

Improving the quality of digital radiographs is one of the most common necessities in medical imaging. One of the most common techniques to improve the quality of a radiographic image that exists in the literature is noise reduction, which is similar to the one being done by Lee *et al*. (2020), where they evaluated the image quality of low-dose digital radiographic images obtained with a new spatial noise reduction algorithm. Other techniques like contrast enhancement exist as well, like the one done by Kushol *et al*. (2019), where we performed contrast enhancement using morphological operators, which could help visualize important bone segments and soft tissues more clearly. Digital radiographic image quality matters not only to the professionals doing the diagnosis but also to the algorithms used as a non-primary way of diagnosing patients, such as deep learning (DL) algorithms. In this study, the investigation on the effects of different radiographic image quality attributes (i.e., contrast, Gaussian blur, Gaussian noise, and salt-and-pepper noise) on the performance of different convolutional neural network (CNN) algorithms has been conducted. The result of this study will be insightful and useful in creating new or improving current techniques to improve the quality of radiographic images, especially x-ray images for DL purposes.

Patients with severe and critical cases of the virus often develop respiratory diseases such as pneumonia, and the need for medical imaging was necessary to determine the severity and the treatment needed for these respiratory diseases. The characteristics of such an infection can be observed by a radiologist and by Deep Learning (DL) methods that can perform deep analysis across radiographic images. This makes early-stage, and precise diagnoses prevent the disease's severity (Hammoudi *et al*., 2020). Subjective evaluation focuses on the perception of quality from the perspective of professionals. Understanding the technology of DL helps researchers and medical practitioners understand the effects of radiographic image quality in computer vision and the value of objective evaluation (Dodge & Karam, 2016). DL uses raw data to automatically discover representations needed for detection and classification. DL models were actually not new in the field of medical imaging; machines that use Computed Tomography (CT) and other medical imaging techniques use Artificial Intelligence (AI) to ensure the quality of medical images produced by these imaging techniques and to enhance the produced alongside other clinical parameters, both Machine Learning (ML) and DL offer fast, automated, and effective strategies to detect and classify abnormalities and extract specific features (Basu *et al*., 2020). ML algorithms where DL falls under have the potential for investment in medicine. From drug discovery to clinical decision-making, the success of ML in recent years can be utilized, especially as medical records are increasingly digitized (Ker *et al*., 2017).

Deep neural networks have the ability to extract sophisticated structures in raw data and, at the same time, hidden features. The amount of data utilized in training the algorithm determines its ability to generalize by ensuring that the data is properly handled, which makes these technologies special and can help remote areas or areas where medical professionals are scarce to provide the needed health care (López-Cabrera *et al*., 2021). The availability of open sources for free access to digitized radiographs from El-Shafai & Abd El-Samie (2020), Kaggle repository, GitHub, Mendeley, and other open sources allows different researchers around the world to make use of these images to create, improve, and develop models that can help and protect society, especially during this pandemic. On the other hand, Basu *et al*. (2020), Jain *et al*. (2021) and Guissous (2019) are some of the studies that utilized open sources in studying and applying DL, specifically Convolutional Neural Network (CNN), in various applications in the medical field, such as the classification of COVID-19 through medical images, skin lesions, and other classification tasks.

The spread of COVID-19 in December 2019 has caused challenges to the healthcare systems worldwide such as containing the virus and preventing the sudden increase in mortality rate. The need for medical and healthcare means a risk for the frontliners in exposing themselves to COVID-19 patients. The virus was devastating, causing millions of people at risk of falling into extreme poverty and necessitating medical attention. This supports the fact that there was a necessity to strengthen COVID-19 testing to isolate and provide treatment for those infected, especially those with severe and critical cases of the virus. There were previous studies during the pandemic that focused on this concern by applying technology to some treatments or diagnoses to limit the interaction of the medical frontliners and patients. Since the data being used in this study are COVID-19 x-ray images (i.e., COVID and Non-COVID), the results of this study could be useful in improving the techniques proposed by recent researchers in literature, particularly in improving the quality and quantity of data being used to predict COVID-19 in radiographic images using

DL, especially CNN algorithms. Although the dataset used in this study is COVID-19 data, the method in this study could be applied as well when it comes to other lung diseases.

This study deals with the lack of understanding in the effects of radiographic image quality attributes in the field of computer vision and deep learning, the challenges to get high quality medical image data for building CNN models and the improvement of healthcare with the use of deep learning. The study aims to investigate the effects of digital radiographic image quality on the performance of CNN models to predict COVID-19 likelihood. Specifically, the study aims to: a) analyze how radiographic image qualities (i.e., contrast, Gaussian blur, Gaussian noise, and salt-and-pepper noise) affect the performance of the CNN models in different performance metrics (i.e., F1 score, AUC, accuracy, precision, and recall); and b) compare the performances of different CNN models in predicting COVID-19. The CNN models to be compared are as follows: DenseNet121, DenseNet169, DenseNet201, Inception-ResNetV2, InceptionV3, ResNet101, ResNet101V2, ResNet152, ResNet152V2, ResNet50, ResNet50V2, VGG16, VGG19, and Xception.

2. METHODOLOGY

2.1 Convolutional Neural Networks (CNNs)

The CNN model sees the input image as an array of matrices. The structure of a CNN model can be divided as follows: the base (i.e., feature learning) and the head (i.e., classification). The base of the CNN is responsible for feature extraction and has three important components, namely: (i) convolution layer, (ii) rectified learning unit (ReLU) activation function, and (iii) pooling layer.

The convolution layer works in a way that it slides a matrix called “kernel” over an input image (or input matrix). The main purpose of sliding the kernel on every region of an input image is to filter out the entire input image, leaving only the important features such as edges. The output image (matrix) of the convolution layer is called the “convolved feature.”

The ReLU activation function will then apply non-linearity in a way that deactivates pixels of an image that have a value less than or equal to zero, leaving only positive values in the matrix. The main purpose of applying ReLU is that it detects the features of images that are filtered by the convolution layer. After that, the output matrix (i.e., convolved feature) from the convolution layer and ReLU will then pass on to the pooling layer.

The pooling layer works in a pretty similar way to the convolution layer. It also includes a small matrix that moves through different regions of an input image (i.e., matrix of convolved features) and selects the maximum (i.e., max pooling) or average (i.e., average pooling) value of pixels in each region. The main purpose of applying a pooling layer is to condense the output image (i.e., the matrix convolved), which is an output of the convolution layer and ReLU activation.

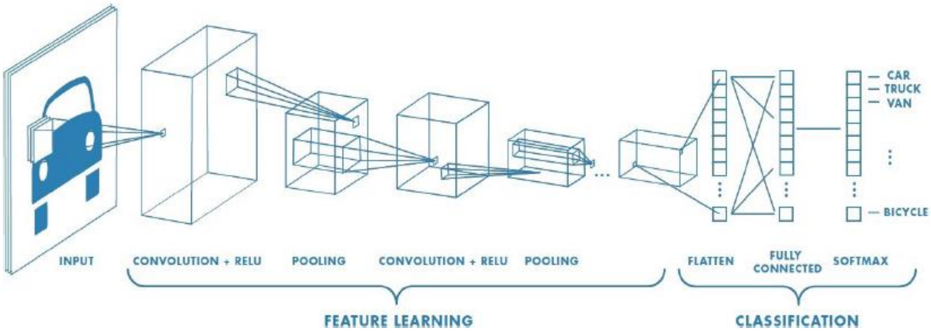


Figure 1. The basic architecture of convolutional neural networks CNN (Rajpal, 2020).

We will discuss some of the general CNN architectures used in building the CNN models in the following subsections.

2.1.1 Visual Geometry Group (VGG) architecture

VGG is a CNN model that was introduced by Simonyan and Zisserman in 2014. This CNN model is one of the best models submitted to ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2014) (Russakovsky *et al*., 2015). This model also surpassed the performance of the first deep CNN model, AlexNet, by using multiple 3 × 3 kernel-sized filters on its convolution layers. The VGG architecture models were trained for several weeks using NVIDIA Titan Black GPUs. Figure 2 shows the standard VGG16 architecture, one of the two most common variants of VGG.

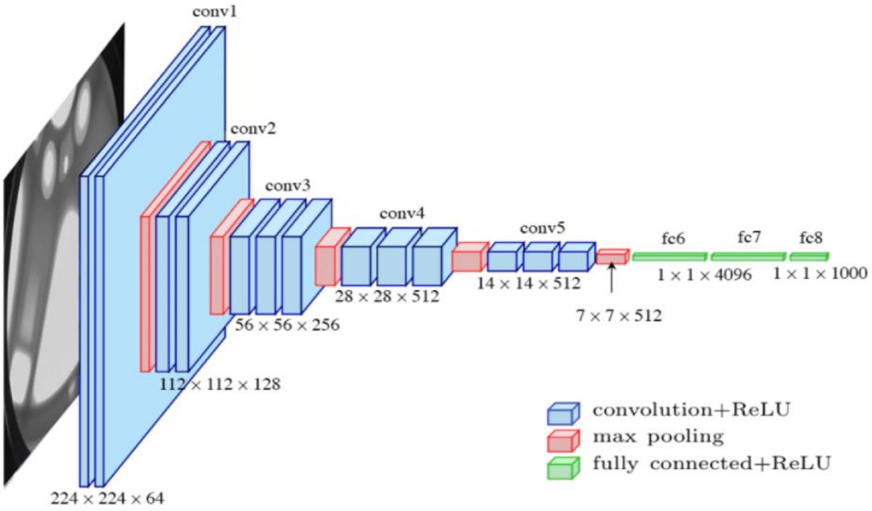


Figure 2. Visualizing VGG16 (Huang *et al*., 2017).

2.1.2 Inception architecture

Inception is a CNN model that was introduced by Szegedy *et al*. in 2015. It focuses on consuming less computational power by modifying the previous Inception architectures. InceptionV3 also demonstrated that high-quality results could be reached with receptive field resolution as low as 79×79 which is helpful in systems for detecting relatively smaller objects (Szegedy *et al*., 2015). A variant of the Inception architecture called InceptionV3 can be visualized in Figure 3.

2.1.3 DenseNet architecture

DenseNet is a CNN model that was introduced by Russakovsky *et al*. in 2016. The advantage of using this model is that it prevents overfitting, leading to false, accurate results, and requires less computation to achieve competitive performance. The CNN models with DenseNet architecture work in a way that the convolved feature matrix that was the product of the convolution layer and ReLU activation function also serves as an input for all other convolution layers. Figure 4 shows the connectivity of the layers in a DenseNet architecture.

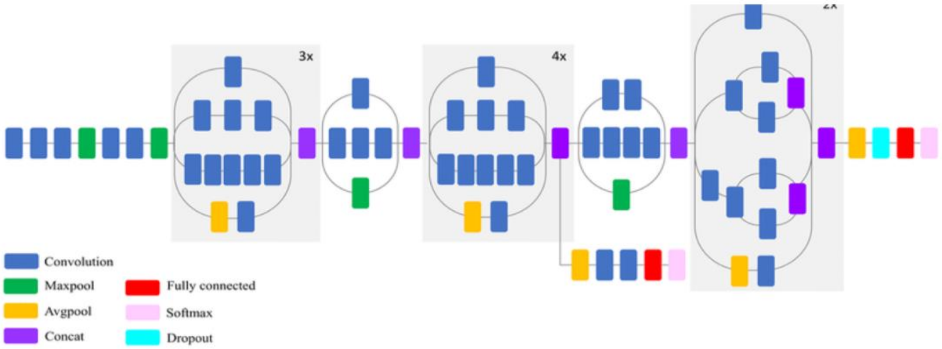


Figure 3. Visualizing InceptionV3 (Mahdianpari *et al*., 2018).

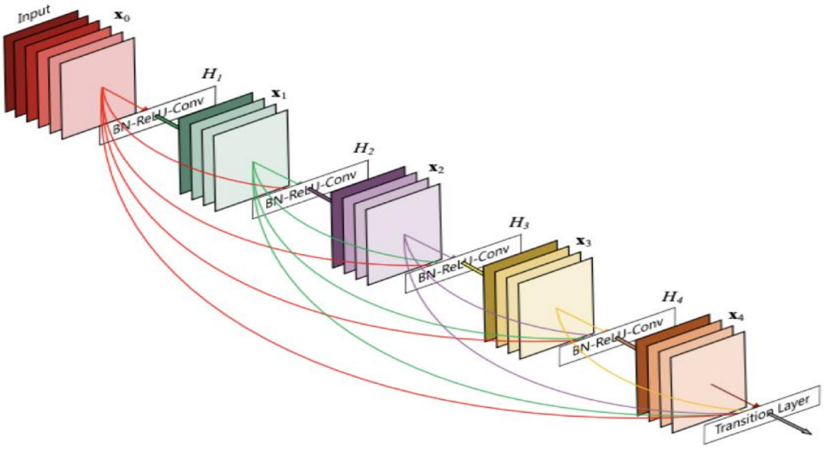


Figure 4. Connectivity of the layers in DenseNet121 (Huang *et al*., 2017).

2.1.4 Residual Network (ResNet) architecture

ResNet is a type of CNN architecture introduced by He *et al*. in 2016. This architecture first introduced the concept of skipping layers. This type of architecture ensures that the higher layers in the model perform as well as the lower layers and not worse. Figure 5 shows what the layers in ResNet50 look like a 50-layer variant of the ResNet architecture.

2.1.5 Xception architecture

Xception is a type of CNN architecture that involves depth-wise separable convolutions. This architecture was first introduced by Francois Chollet, who was also the founder of Keras (Chollet, 2017). In addition, Xception is an improved version of Inception. Figure 6 shows the connectivity of layers in the Xception architecture.

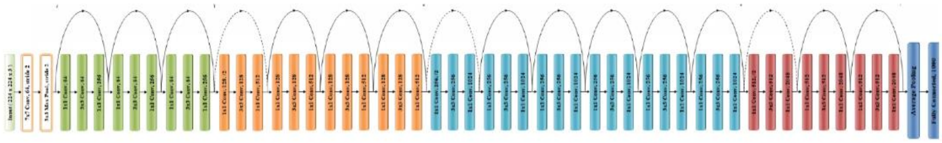


Figure 5. Visualizing the ResNet50 (Sachan, 2017).

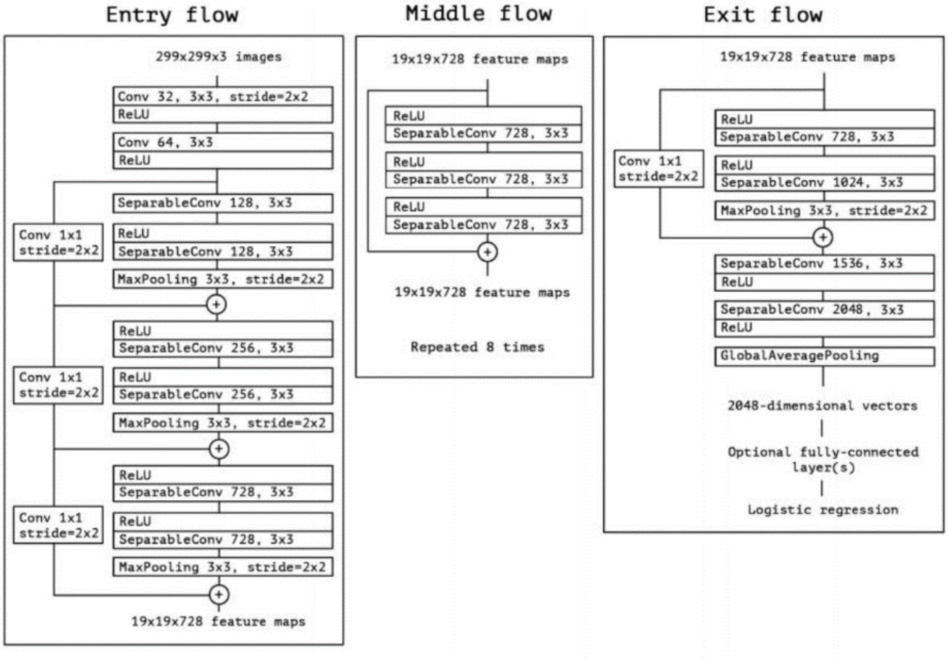


Figure 6. Diagram of Xception architecture (Chollet, 2017).

2.2 Data collection and preparation

We used the dataset of El-Shafai and El-Samie (2020) as a training set. The publishers of the dataset applied different augmentation techniques to generate about 17,099 x-ray and CT images combined. The dataset contains two main folders: one for the x-ray scans, which includes two separate subfolders of 5,500 non-COVID images and 4,044 COVID images. The other folder contains the CT images. It includes subfolders of 2,628 non-COVID images and 5,427 COVID images. In this study, we only used the x-ray scans, which have a total of 9,544 images, for training the transfer learning models.

Meanwhile, we used the dataset from Kaggle, published by Amanullah Asraf in 2020, as the validation and test set. The said dataset was gathered from different sources. The 613 x-ray images of COVID-19 cases were collected from the combined datasets of COVID-19 Image Data Collection (Cohen *et al*., 2020), Actualmed COVID-19 Chest X-ray Dataset Initiative (Wang *et al*., 2020), The Cancer Imaging Archive (TCIA), and the Italian Society of Radiology (SIRM) (Cohen *et al*., 2020). The publisher also used 912 already augmented x-ray images from the Augmented COVID-19 X-ray Images Dataset (Alqudah, 2020). In addition, the said dataset also contains 1,525 images of pneumonia cases and 1.525 x-ray images of normal cases which were collected from the Kaggle repository of a published article (Kermany *et al*., 2018) and the National Institutes of Health (NIH) dataset (Wang *et al*., 2017). The authors of this paper only used the x-ray images of COVID-19 and normal cases. After gathering the datasets from their respective sources, the images from each dataset were resized to 224×224 resolution using OpenCV (Bradski & Kaehler, 2000), which is required for pre-trained model inputs, and split into training, validation, and test sets as shown in Table 1.

2.3 Radiographic image quality attributes

The radiographic image quality of the images from the dataset was manipulated in ten levels of the following attributes: contrast, Gaussian blur, Gaussian noise, and salt-and-pepper noise. The manipulation of contrast was based on the proposed method of Haeberli and Voorhies (1994) called the interpolation and extrapolation. Applying interpolation reduces the contrast while extrapolation increases it. The authors used a blending factor of 0 to 1 with a step of 0.1. A blending factor of 1 returns the original radiographic image quality and blending factors less than 1 reduce the contrast. Figure 7 shows an illustration of decreasing and increasing contrast using the concept of interpolation and extrapolation.

Table 1. Dataset split ratio and its percentage.

Set	COVID	Non-COVID	Total
Train Set	4044	5500	9544 (75.78%)
Valid Set	763	763	1526 (12.12%)
Test Set	762	762	1524 (12.10%)
Total	5569 (44.22%)	7025 (55.78%)	12594 (100%)

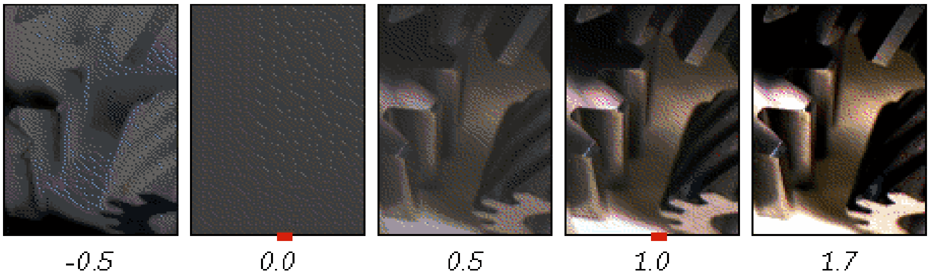


Figure 7. Illustration of contrast interpolation and extrapolation, the values refer to the contrast factor which is a unitless value (Haeberli & Voorhies, 1994).

The blurring of an image was done using Gaussian blur. It works in a way that it slides a matrix called a "kernel" (smaller than the matrix of the input image), applies weights to each value in the sliding kernel (the values close to the center of the matrix are given more weights than those far away) and computes the average of all pixel values in a kernel (Carpentries, 2021). The computed average value in the kernel will be used to replace the old pixel. We varied the standard variation of the Gaussian from 1 to 10 in steps of 1. The size of the kernel windows is set to 4 times the standard deviation. The Python library called OpenCV was used to implement Gaussian noise and Gaussian blur. The third attribute is the Gaussian Noise which has a probability density function equal to Gaussian distribution. This type of noise disturbs the gray values in digital images (Boyat & Joshi, 2015). It is also called "electronic noise" because it arises in amplifiers or detectors (Swain, 2018). The magnitude of Gaussian noise depends on and is proportional to the standard deviation σ of Gaussian distribution which can be expressed as

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(z-\mu)^2}{2\sigma^2}} \tag{1}$$

where z is the gray level, μ is the mean of average value of z and σ is the standard deviation. We varied the standard deviation of the noise from 10 to 100 with steps of 10. Finally, the third radiographic image quality is the salt-and-pepper noise which uses the Probability Density Function (PDF) to randomly distribute the light and dark color pixels on a given image. The PDF of salt-and-pepper noise can be defined as

$$p(z) = \begin{cases} p_a & \text{for } z = a \\ p_b & \text{for } z = b \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where p_a, p_b are the probability density functions (PDFs), $p(z)$ is the distribution of salt-and-pepper noise in an image and a, b are values between 0 and z. We applied the mentioned attributes to radiographic images in ten levels which can be seen below in Figure 8.

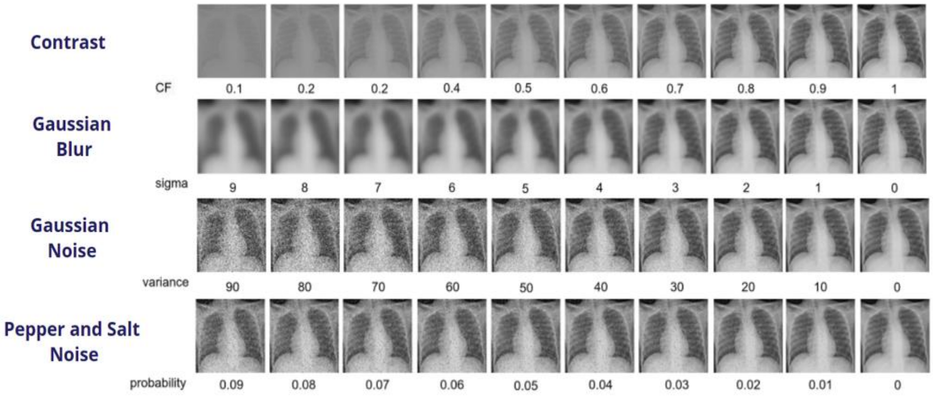
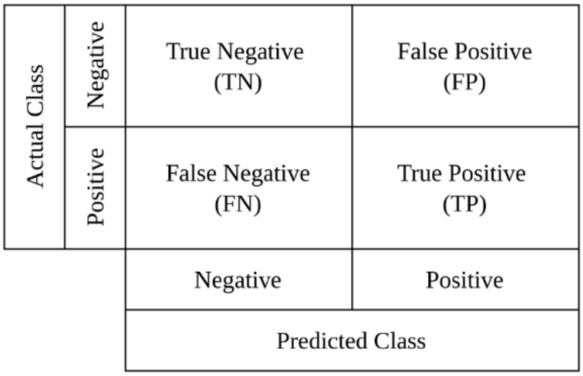


Figure 8. Illustration of different levels of contrast, Gaussian blur, Gaussian noise, and salt-and-pepper noise on an x-ray image.

2.4 *Performance metrics*

Confusion matrix: A confusion matrix is a type of metric to visualize the quantity of samples that are correctly or wrongly classified or predicted by the model during prediction. Figure 9 shows an example of a confusion matrix where a predicted sample can be categorized as True Positive (TP), False Negative (FN), False Positive (FP), and/or True Negative (TN). The y-axis refers to the actual samples, while the x-axis refers to the predicted samples. When it comes to medical images, the term “Positive” here refers to the samples (i.e., images) with an infected disease, while the term “Negative” refers to the samples without the disease. In this study, TP refers to the samples with confirmed COVID cases, which were also correctly predicted by the models to have COVID. FN refers to the samples with confirmed COVID cases but wrongly predicted by the models as Non-COVID. FP refers to the Non-COVID samples but wrongly predicted by the models to have COVID. TN refers to the Non-COVID samples, which were also correctly predicted by the models as Non-COVID.



Actual Class	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)
		Negative	Positive
		Predicted Class	

Figure 9. Illustration of confusion matrix.

Precision and recall: When it comes to binary classification problem, precision refers to the number of true positives (TP) divided by the total number of positive predictions (i.e., TP + FP). Increasing precision reduces the number of FP. Meanwhile, recall refers to the number of TP divided by the total number of TP and FN. Maximizing the recall will minimize the FN. Here is the formula to calculate the precision and recall of the model respectively:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP+FN} \tag{3}$$

F1-score: The performance metric F1-score is the weighted average of precision and recall. It takes into account the FP and FN samples, which makes it useful when dealing with imbalanced datasets. The best score it can give is 1, indicating that the model correctly predicted all the respective labels of samples, while 0 is the worst, as it indicates that the model incorrectly predicted all the respective labels of samples. To calculate the F1-Score, one can use the following formula:

$$F1\text{-}Score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right) \tag{4}$$

Area under the curve of receiver operating characteristic (AUC-ROC): The AUC-ROC (or commonly called AUC) performance metric is one of the most common and important metrics for classification problems. To better understand the AUC, one must first understand how ROC works. An ROC curve is a curve of probability that determines if the model is able to separate the samples into positive and negative classes. The green and red curves below (Figure 10) are the distributions of TN and TP, respectively. The ideal scenario here is that the model can perfectly separate the classes (i.e., TN and TP). Meanwhile, the worst scenario is that the model was not able to distinguish the difference between classes. To plot the ROC (i.e., the colored orange curve in Figure 10), one must compute the True Positive Rate (TPR) and False Positive Rate (FPR) with many different thresholds. TPR is just another name to describe recall. Meanwhile, FPR can be computed using the ratio between FP and FP+TN. The area under the ROC curve is what the AUC-ROC is all about. If the model was able to distinguish the difference between classes perfectly, it would give a value of 1, which is the highest value possible. The worst value it can give is 0.5, which means that the model was not able to distinguish the difference between classes. There are also cases where AUC becomes 0, which happens when the model wrongly classifieds the positive class as negative or vice versa.

2.5 Software

The Python library called PIL (Clark, 2021) was used to manipulate images at different levels of radiographic image quality attributes, specifically contrast and salt-and-pepper noise. Meanwhile, another Python library called OpenCV (Bradski & Kaehler, 2000) was used to apply different levels of Gaussian noise and Gaussian blur to the images. The library PIL is mainly used for image processing and manipulation in Python, and OpenCV, on the other hand, is used not only for image processing but also for other artificial intelligence applications in general.

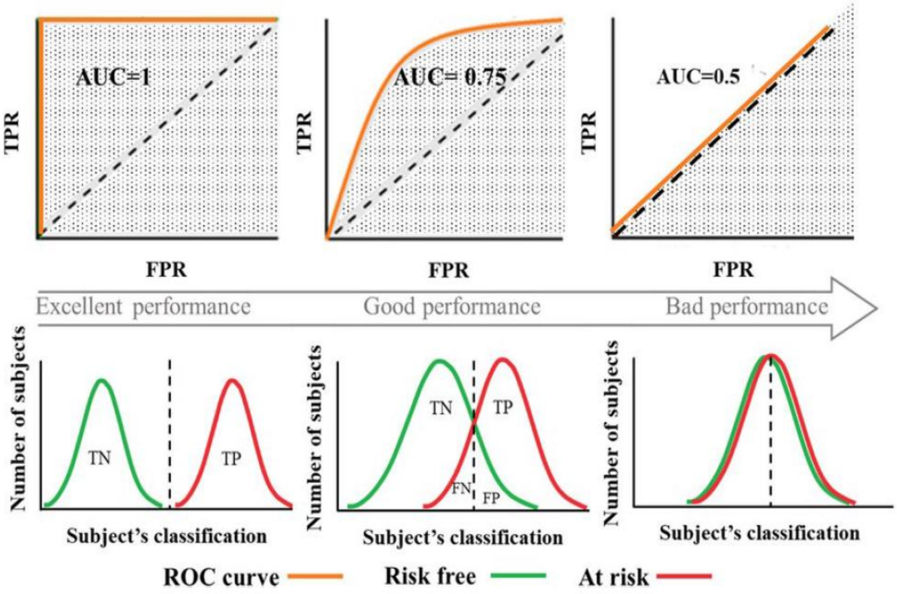


Figure 10. Illustration of area under the curve (AUC).

2.6 *Schema of the experimental design*

Figure 11 shows the detailed schematic diagram of the methodology, from gathering the data to conducting the experiment. In summary, data was gathered from various sources; primarily Mendeley Data and Kaggle, with data gathered from Kaggle coming from various sources as well (refer to Section 2.2). The data treatment is where the data has been resized in order to be compatible with the imported deep transfer learning models. Besides, this is also where the data has been split into three sets (i.e., the training set, the validation set, and the test set). After that, model training was conducted using the data from the training set and validation set. Meanwhile, different radiographic image quality manipulations were applied to the test set, namely applying contrast, Gaussian noise, Gaussian blur, and salt-and-pepper noise. This results in having different kinds of test sets, namely the test set with original radiographic image quality and with different levels of radiographic image qualities (i.e., with applied contrast, Gaussian noise, Gaussian blur, and salt-and-pepper noise, respectively). During test evaluation, the trained models have been tested against images with original and manipulated radiographic image qualities. With that, there are five types of results generated from the experiment, namely the results for original radiographic image quality and manipulated radiographic image quality (i.e., contrast, Gaussian noise, Gaussian blur, and salt-and-pepper noise).

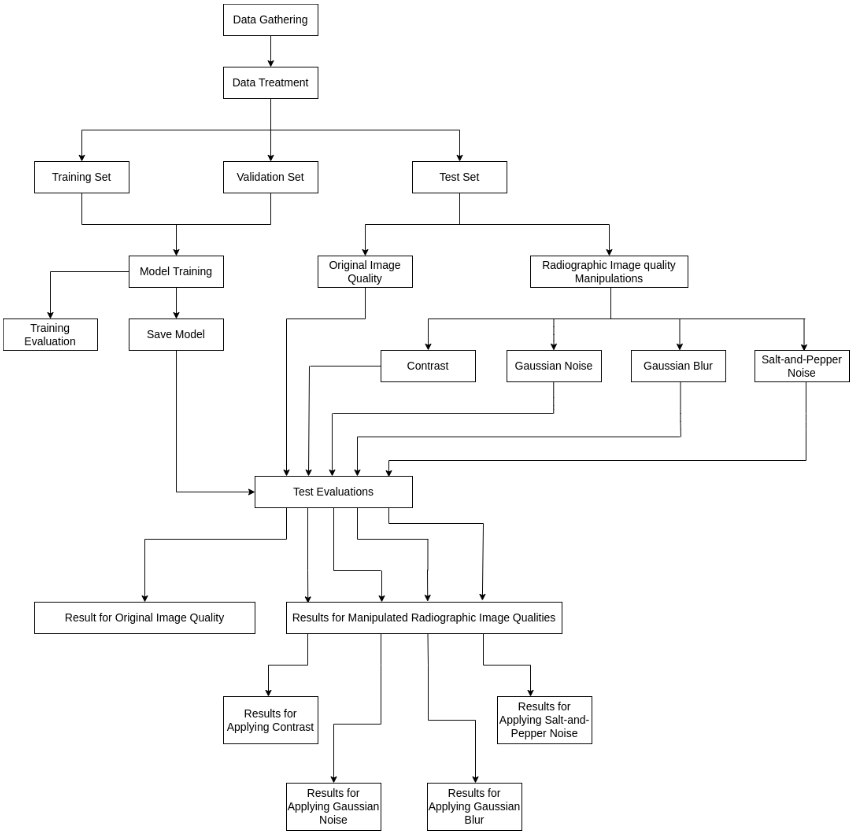

Data Gathering
Data Treatment
Training Set
Validation Set
Test Set
Model Training
Original Image Quality
Radiographic Image quality Manipulations
Training Evaluation
Save Model
Contrast
Gaussian Noise
Gaussian Blur
Salt-and-Pepper Noise
Test Evaluations
Result for Original Image Quality
Results for Manipulated Radiographic Image Qualities
Results for Applying Contrast
Results for Applying Salt-and-Pepper Noise
Results for Applying Gaussian Noise
Results for Applying Gaussian Blur

Figure 11. Schema of the experiment design.

3. RESULTS AND DISCUSSION

3.1 Performance of models on original radiographic image quality

Figure 12 shows the performances of the models against the original radiographic image quality expressed in F1 score. This result is also the performance of the models when classifying COVID-19 in radiographic images. The family of transfer learning models with a DenseNet architecture outperformed other experimental models like Inception, ResNet, VGG, and Xception models in terms of F1 score. Specifically, Densenet201 recorded the highest performance with an F1 score of 0.969, followed by DenseNet121 and DenseNet169 with F1 scores of 0.9579 and 0.948, respectively. Note that the primary metric to determine the best performing model in this study is the F1 score, as it takes into account both false positives and false negatives. Other performance metrics were used as well to relate the performances of the models with the results of other studies, like the use of AUC, accuracy, recall, and precision.

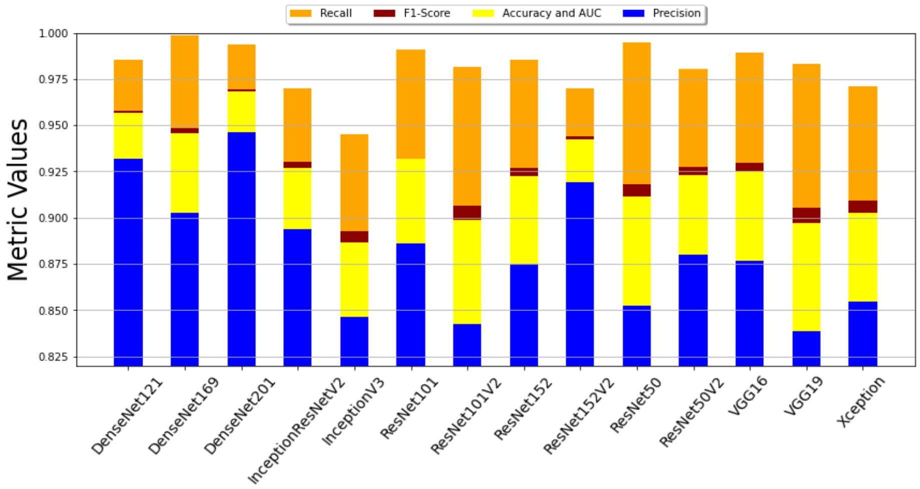


Figure 12. Performance of the models against the original image quality.

In Figure 12, accuracy and AUC were unified as they have almost the same values for all the models. As shown in Figure 12, DenseNet also achieved the highest performance in terms of other metrics, with an AUC = 0.9685, accuracy = 0.9685, precision = 0.946, and recall = 0.993. Table 2 shows a more comprehensive result about the performance of each CNN model against the original radiographic image quality.

3.2 Effects of decreasing contrast

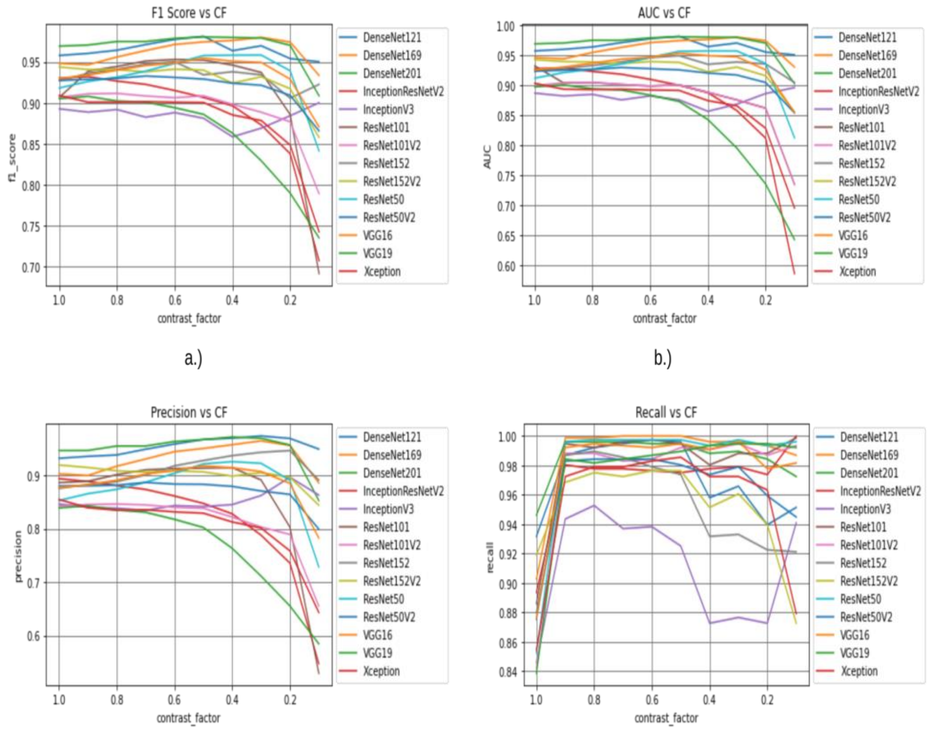
In the field of medical imaging, most of the time, contrast plays a role in improving radiographic or CT images. This is to improve the visibility of some organs and make them easier to detangle for a better diagnosis. Figure 13 shows the effects of the contrast on the performance of DL models in terms of F1-score, AUC, precision, and recall. While the performance of the models slowly decreases through the decrease of the contrast factor, there is a slight increase which later on decreases rapidly. The effect of decreasing contrast in model performance indicates the models' capability to decrease the number of false positive and false negative samples. Meanwhile, the result in AUC shows the models' capability to separate the positive from negative classes (i.e., COVID and Non-COVID). Decreasing the contrast factor contrast also decreases the AUC of the model. Meanwhile, precision can be defined as the quality of the positive prediction made by the model. The higher the precision is, the higher the quality of the detected COVID-19.

Table 2. Performance of the models against the original radiographic image quality expressed in different metrics (F1 score, AUC, accuracy, precision, and recall).

Model	F1	AUC	Accuracy	Precision	Recall
DenseNet121	0.957	0.956	0.956	0.931	0.985
DenseNet169	0.948	0.945	0.945	0.902	0.998
DenseNet201	0.969	0.968	0.968	0.946	0.993
InceptionResNetV2	0.930	0.927	0.927	0.893	0.969
InceptionV3	0.892	0.886	0.886	0.846	0.944
ResNet101	0.906	0.931	0.931	0.886	0.990
ResNet101V2	0.906	0.898	0.898	0.842	0.981
ResNet152	0.927	0.922	0.922	0.875	0.985
ResNet152V2	0.943	0.942	0.942	0.919	0.969
ResNet50	0.918	0.911	0.911	0.852	0.994
ResNet50V2	0.927	0.923	0.923	0.879	0.980
VGG16	0.929	0.925	0.925	0.876	0.989
VGG19	0.905	0.896	0.896	0.838	0.982
Xception	0.909	0.902	0.902	0.854	0.971

3.3 Effect of increasing Gaussian blur

In the field of medical imaging, Gaussian blur is commonly used in many noise reduction methods, which play a role in improving the readability of a radiographic image. The variable used to manipulate Gaussian blur is the standard deviation (i.e., sigma). Gaussian blur decreases the performance of the models in all metrics more rapidly than that of contrast. The effect of Gaussian blur in F1-score shows a more congregated trend in the first half than any other metric. This concludes that Gaussian blur could affect the performance of the models when they are slightly close to one another, while the second half of the plot (Figure 14a) shows a sparser trend in model performance in F1-score. The trends in the performance of the models in terms of F1-score indicate the capability of the models to minimize the quantity of false positive and false negative samples. Meanwhile, the trends in model performance in AUC and precision are slightly identical, and they both decrease more rapidly with increasing Gaussian blur. The plots of models' performance in AUC and precision shown in Figures 14b and 14c indicate the models' capability to separate the positive from negative classes (i.e., COVID and Non-COVID) and the quality of the positive predictions by the models, respectively.



a.)

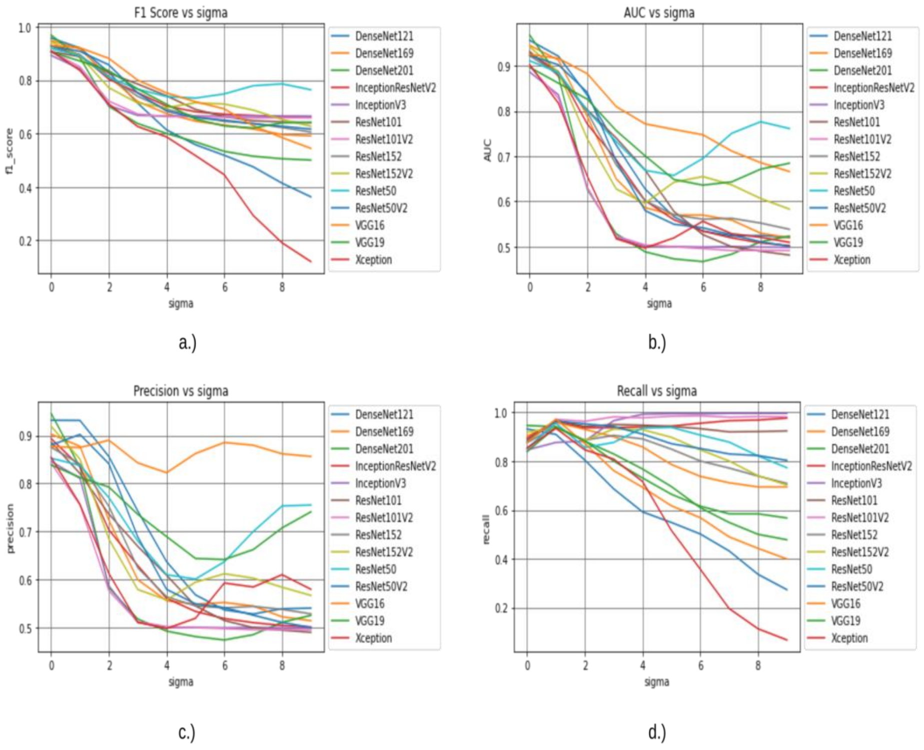
b.)

c.)

d.)

Figure 13. Effects of contrast on the performance of CNN models measured using (a) F1 score, (b) AUC, (c) Precision, and (d) Recall by manipulating the values of contrast factor in the image dataset.

Finally, the plot on recall (figure 14d) shows a mix of trends between models. Some models show a rapid increase in recall, some of which converge to 1.00, while others decrease rapidly as the gaussian blur increases. The plot of models' performance in recall indicates the models' ability to make positive predictions under the increasing levels of gaussian blur. In medical imaging, applying gaussian blur or blur, in general, is commonly used in medical image enhancement as a noise reduction method. Compared to models' performance against decreasing contrast (refer to section 3.2), experiment models are more susceptible to gaussian blur. Also, the performance of dl models against the decreasing gaussian blur is different from the result of dodge and karam (2016) where the performance of dl models shows no sign of an increase in performance. Note that all of the stated results in this study about the effects of gaussian blur could be biased in favor of deep learning algorithms, and the result might be different if other applications are involved.



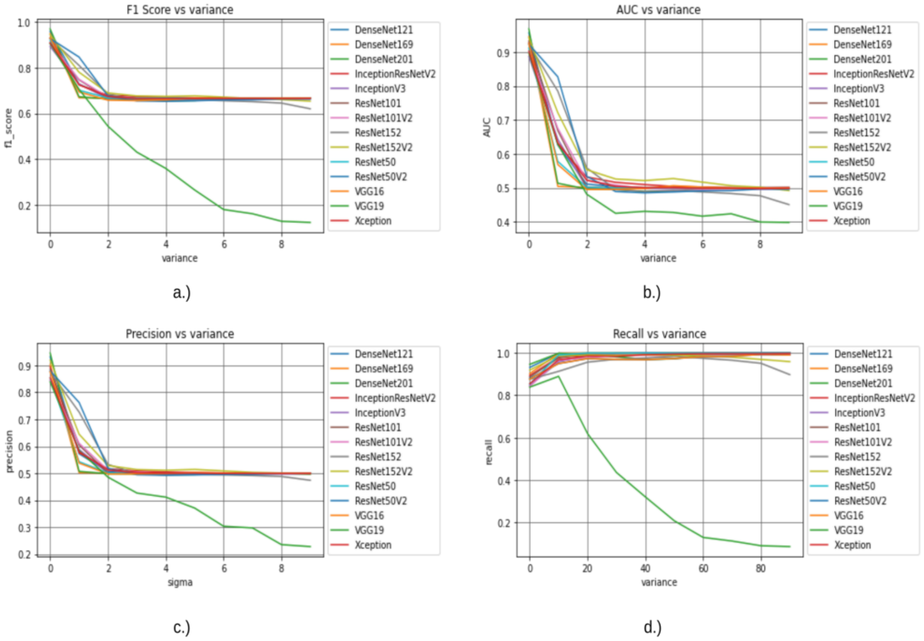
a.) b.)

c.) d.)

Figure 14. Effects of Gaussian blur on the performance of CNN models measured using (a) F1 score, (b) AUC, (c) Precision, and (d) Recall by manipulating the values of sigma in the image dataset.

3.4 *Effects of Gaussian noise*

Gaussian noise is a type of noise that commonly occurs during medical image acquisition. Specifically, it occurs due to camera or device inefficiency being used to capture a medical image. Gaussian noise has never been helpful when diagnosing, and so, most of the time, the existing studies in literature focus on removing not only Gaussian noise but noise in general. The results present the effect of Gaussian noise on the performance of different DL models expressed in a different matrix. It is visible that, for all metrics, Gaussian noise rapidly reduced the performance of the DL models.



a.) b.)

c.) d.)

Figure 15. Effects of Gaussian noise on the performance of CNN models measured using (a) F1 score, (b) AUC, (c) Precision, and (d) Recall by manipulating the values of variance in the image dataset.

In terms of F1-score, the performance of DL models converged at 0.65 – 0.66. However, there is still a trend that differs from the others, like the trend on the performance of the model VGG19, which continuously decreases even in other metrics. The trend in F1-score indicates the capability of the DL models to reduce false positive and false negative predictions. In terms of AUC and Precision, the performance of DL models converged in 0.5 after variance = 2, which means the models were already giving random predictions at this point. Finally, the performance of DL models in terms of recall shows a sudden increase in trend. Recall converges up to 1.0 when increasing the Gaussian noise. That is, increasing Gaussian noise increases the ability of the model to predict the positive class (i.e., COVID), which is the same when reducing contrast. These results are somewhat different from the results of Dodge and Karam (2016) where they also evaluated the effects of Gaussian noise in non-medical images. In their study, although the performance of the DL models is decreasing, the trends are not decreasing rapidly, which is different from the results found in this study (Figure 15).

3.5 *Effects of salt-and-pepper noise*

Salt-and-pepper noise generally occurs when there is an error during analog-to-digital conversion. It can be visualized as random black and white pixels in an image containing the noise. This type of noise, like Gaussian noise, has never been useful in radiographic image diagnosis. In terms of F1-score, the performance of DL models against increasing salt-and-pepper noise rapidly decreased and converged to approximately 0.66, which is almost the same as Gaussian noise.

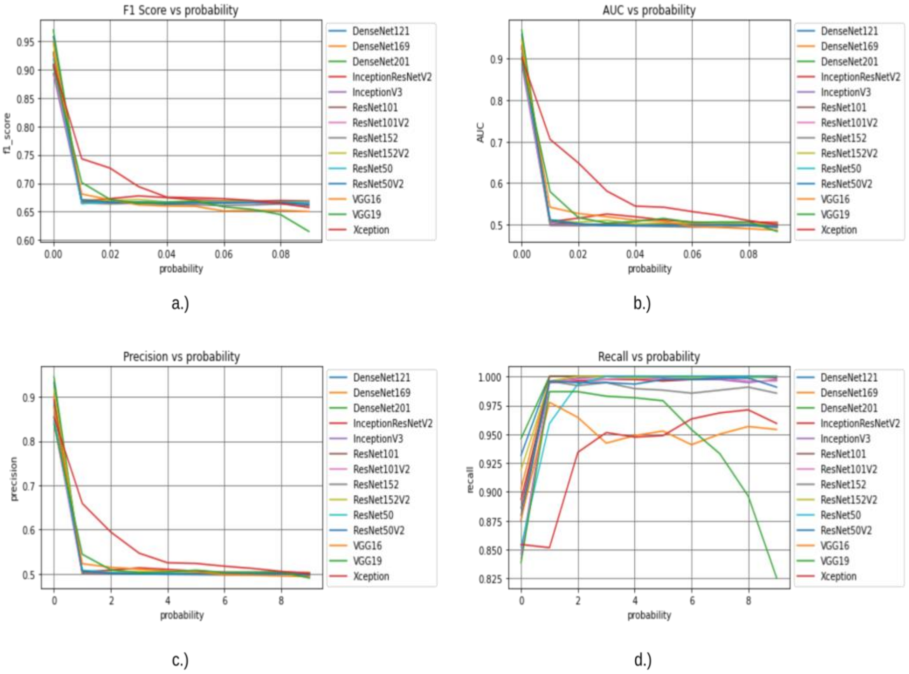


Figure 16. Effects of salt-and-pepper noise on the performance of CNN models measured using (a) F1 score, (b) AUC, (c) Precision, and (d) Recall by manipulating the values of probability in the image dataset.

The trends in Figure 16a indicate the capability of the models to reduce false positive and false negative predictions. Meanwhile, in terms of AUC and precision, the performance of the models also rapidly decreased and converged to 0.5, which is also the same as in Gaussian noise. The AUC and precision of 0.5 only demonstrate that the models are already making random predictions. On the other hand, recall quickly increases for the majority of the models. This indicates that the capability of the models to predict the positive class (i.e., COVID) increases as the salt-and-pepper noise also increases.

4. CONCLUSIONS

Investigation of the effects of digital radiographic image quality on the performance of DL models can lead to other research milestones, especially in improving the performance of DL models in the future. This is because the outcomes of such an investigation could be useful in improving the quality of data being used to train DL models. Through the use of insightful results from this study, techniques such as data augmentation and medical image quality enhancement could be implemented more efficiently. Since the dataset used in this study focuses on identifying COVID-19 cases, this study is also useful when improving current DL models for COVID-19 detection existing in the literature, specifically by improving the quality and quantity of data. While the data being used is focused on COVID-19 detection, the results of this study could be useful and applied as well when dealing with other digital radiographic images. In this study, different levels of digital radiographic image qualities such as contrast, Gaussian blur, Gaussian noise, and salt-and-pepper noise were found to have a significant effect on the performance of experimented DL models. The DL models were found to be resilient at decreasing levels of contrast. In fact, the models' performance was also found to be improving at some point before it finally declined at approximately contrast factor = 0.3. Also, DL models are less resilient to increasing Gaussian blur compared to contrast. All of the models have shown no resiliency to increasing noise (i.e., Gaussian noise and salt-and-pepper noise). Another insight worth noting is that the recall of the DL models was significantly increased during the manipulation of the said attributes, causing the models to be biased in favor of samples with COVID cases. The insights from the results of this study could be used as a basis to improve the current performance of several DL models, particularly CNN, by improving the quality and quantity of data, which is more of a data-centric approach than a model-centric one.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge the institutions who funded this research, Polytechnic University of the Philippines Research Institute of Science and Technology (PUP-RIST) and Department of Science and Technology - Science Education Institute (DOST-SEI).

6. REFERENCES

Alqudah, A. M. (2020). Augmented COVID-19 x-ray images dataset.

Basu, S., Mitra, S., & Saha, N. (2020). Deep learning for screening covid-19 using chest x-ray images. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 2521-2527). IEEE.

Boyat, A. K., & Joshi, B. K. (2015). A review paper: Noise models in digital image processing. *arXiv preprint arXiv:1505.03489*.

Bradski, G., & Kaehler, A. (2000). OpenCV. *Dr. Dobb's Journal of Software Tools*, *3*(2).

Carpentries, T. (2021). Blurring images. https://datacarpentry.org/image-processing/06-blurring/

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).

Clark, A. (2021). Pillow 8.2.0. https://pypi.org/project/split-folders/#description

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020). Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*.

Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. In *2016 Eighth international conference on quality of multimedia experience (QoMEX)* (pp. 1-6). IEEE.

El-Shafai, W., & Abd El-Samie, F. (2020). Extensive COVID-19 x-ray and CT chest images dataset. *Mendeley Data*, *3*(10).

Guissous, A. E. (2019). Skin lesion classification using deep neural network. *arXiv preprint arXiv:1911.07817*.

Haeberli, P., & Voorhies, D. (1994). Image processing by linear interpolation and extrapolation. *IRIS Universe Magazine*, *28,* 8-9.

Hammoudi, K., Benhabiles, H., Melkemi, M., Dornaika, F., Arganda-Carreras, I., Collard, D., & Scherpereel, A. (2021). Deep learning on chest x-ray images to detect and evaluate pneumonia cases at the era of COVID-19. *Journal of Medical Systems*, *45*(7), 75.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

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

Jain, R., Gupta, M., Taneja, S., & Hemanth, D. J. (2021). Deep learning based detection and analysis of COVID-19 on chest x-ray images. *Applied Intelligence*, *51,* 1690-1700.

Ker, J., Wang, L., Rao, J., & Lim, T. (2017). Deep learning applications in medical image analysis. *IEEE Access*, *6*, 9375-9389.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122-1131.

Kushol, R., Raihan, M., Salekin, M. S., & Rahman, A. B. M. (2019). Contrast enhancement of medical x-ray image using morphological operators with optimal structuring element. *arXiv preprint arXiv:1905.08545*.

Lee, W., Lee, S., Chong, S., Lee, K., Lee, J., Choi, J. C., & Lim, C. (2020). Radiation dose reduction and improvement of image quality in digital chest radiography by new spatial noise reduction algorithm. *Plos One*, *15*(2), e0228609.

López-Cabrera, J. D., Orozco-Morales, R., Portal-Diaz, J. A., Lovelle-Enríquez, O., & Pérez-Díaz, M. (2021). Current limitations to identify COVID-19 using artificial intelligence with chest x-ray imaging. *Health and Technology*, *11*(2), 411-424.

Mahdianpari, M., Salehi, B., Rezaee, M., Mohammadimanesh, F., & Zhang, Y. (2018). Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sensing*, *10*(7), 1119.

Rajpal, G. (2020). A comprehensive guide to convolution neural network. https://medium.com/swlh/a-comprehensive-guide-to-convolution-neural-network-86f931e55679

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115, 211.*

Sachan, A. (2017). Detailed guide to understand and implement ResNets. *URL:* https://cvtricks.com/keras/understand-implement-resnets/.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Swain, A. (2018). Noise in digital image processing. https://medium.com/image-vision/noise-in-digital-image-processing-55357c9fab71

Szegedy, C., Vanhoucke, V., Loffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

Wang, L., Wong, A., Lin, Z. Q., McInnis, P., Chung, A., Gunraj, H., Lee, J., Ross, M., VanBerlo, B., Ebadi, A., & Al-Haimi, A. (2020). Actualmed COVID-19 chest x-ray dataset initiative. *URL:* https://github.com/agchung/Actualmed-COVID-chestxraydataset

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097-2106).