Automated Detection of COVID-19 using Convolutional Neural Networks and Generative Adversarial Networks

Ultan Kearns
Department of Computing
ATU Letterkenny

I. Abstract

The limited amount of data available for training CNNs to recognize COVID-positive and COVID-negative patients has led to a number of researchers training their models on mislabelled or imbalanced datasets. The limited number of robust models in the field has also led to a number of commercial diagnostic models which could be potentially harmful. In this paper we show that there is some promise to synthetically augmenting data through the use of generative deep learning. Throughout the research we use a variety of different methods to generate the new synthetic data and to balance the dataset by augmenting minority classes, to bring them in balance with the majority classes. The results of this research have shown that the use of synthetic augmentation can improve the accuracy of some CNN models. Although the CNN models were unable to surpass the currently achieved validation set accuracy of the COVID models mentioned in the literature review(the majority of these models attained a validation set accuracy of $> 98\%$)[1][2] it is worth nothing that one of the top models(Extensive CNN CT EfficientNetV2S) trained achieved a test set accuracy of 96.97% and a test set loss of 0.1579 when trained on a larger dataset which contained 4,655 more images than the model mentioned in the literature review. Similar results were also achieved when training models on the radiography dataset, the top model(EfficientNetV2S) achieved a test set accuracy of 96% with a loss of 0.1056 when trained on a much larger dataset containing 60,933. It is also worth noting that this research was conducted using a test set whereas all of the relevant existing models evaluated in the literature review only used a validation set, which indicates that the researchers may have overfit the validation set. When evaluating the results of this research it is important to remember that a larger dataset size will increase the model's ability to generalize and using a test set to evaluate the performance provides us with a much less biased analysis of the model's accuracy.

II. Introduction

The crux of the research conducted centred around the question "can generative deep learning be used to improve the results of currently existing models used to automate the diagnosis of COVID-19?" When beginning the study we researched various different architectures of GANs and various architectures of CNNs as well as existing paradigms used by other researchers when designing and training these models. The main GAN architecture used in this study is the Deep Convolutional GAN or DCGAN for short. This was the GAN which had the most success in generating synthetic COVID-19 X-rays, CT scans, and masks. Variational Autoencoders (VAEs) were also used but had limited success when augmenting the datasets. Transfer learning was also used with a number of different pre-trained model architectures all trained using the ImageNet dataset for the use of automating COVID-19 diagnosis and many were able to outperform the the non transfer learning models. The research goal was achieved with many of the models trained on the augmented datasets outperforming the models trained on the non-augmented datasets.
The research conducted used a number of datasets to train both the CNNs and the GANs. Three datasets were used to train the GAN models: the COVID-19 Radiography Dataset, the Extensive COVID-19 dataset, and the COVID-19 X-Ray dataset. There was also a fourth dataset used to train CNN models: the COVID-19 Chest X-Ray dataset. However, this dataset was not suitable to train GAN models, as there are 11 classes and some classes are extremely underrepresented. Thus in an effort to conserve resources, the decision was made not to use this dataset when training GANs. The CNN models for this dataset were still documented as the effect of transfer learning was still worthy of investigation when using a severely limited dataset.

III. Literature Review

A number of papers were reviewed during the course of this research to determine current paradigms in generative deep learning and computer vision. This section began with researching and analysing currently existing models which were used in the automated detection of COVID-19 in patients. In a paper by Mahmoudi, Benamour et al[2]. Researchers investigated deep learning approaches to aid in the creation of an automated diagnostic tool using computed topography scans. The researchers found that segmenting the infected region of the patients scan and using a technique called Contrast Limited Adaptive Histogram Equalization was beneficial to creating a homogenous dataset and aiding in the improvement of the model's performance. The researchers removed black slices from the images so that only the region of interest was shown

and used a U-net architecture for more timely and accurate segmentation of images. Four-fold cross-validation was also used in the training of this model. The model achieved an accuracy of 98% when diagnosing patients. Despite the high accuracy the model suffered from lack of data, as only 20 CT Scans were used to train and validate the model (40 images in total). The model was also deprived of a test set which may have inflated the accuracy as it is highly likely the model was overfitting to the validation set. Due to the lack of data and the high cost associated with data collection the researchers were unable to collect more data for a test set to provide a less biased and more objective result.

There is also a high potential for bias in this study as the limited amount of data and possibly mislabelled or misclassified data may have had an adverse effect on analyzing the results. This links back to a term "Frankenstein Datasets" referring to datasets which are mislabelled, suffer from lack of data-quality, biased, and spliced together from different sources. Frankenstein datasets were used initially in the early stages of the pandemic as many researchers and companies rushed to find a solution to training a high quality automated COVID-19 diagnostic tool to alleviate the heavy burden placed upon medical staff at the time.

The issue with limited data and imbalance between classes within the datasets is shown in the next study analyzed, seems to be a recurring theme when analyzing currently existing automated COVID-19 diagnostic tools. In a paper by Islam, Islam, and Asraf [1] the researchers used LSTM (Long Short Term Memory) to automate the diagnosis of COVID-19. The model trained as part of this research used a much larger dataset than the study analyzed previously [2]. The dataset used contained 4,575 X-ray images in total with 1,525 images being COVID-19 Positive. The model achieved an area under curve accuracy of 99.4%, an accuracy of 99%, a specificity of 99.2%, and a sensitivity of 99.3%. The researchers suggest that the study may have been improved if more data were available to them. The lack of data is clearly visible from the size of the dataset and the imbalance between classes. The model does appear to improve upon the initial model mentioned in the previous study [2] as it is trained on a far larger dataset and would have an increased ability to generalize. As shown when comparing the accuracy of this model with the one mentioned previously, this model has a much higher accuracy.

The model trained as part of this study appeared to greatly benefit from the use of an LSTM architecture. The use of the model architecture allows the neural network to remember and carry previously learned information to layers deeper in the network, greatly helping with the problems of vanishing and exploding gradients[3].

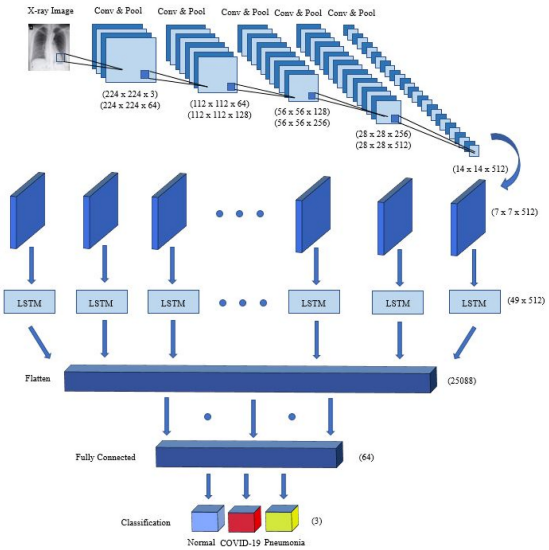

Fig. 1: Overview of a Typical LSTM Network [1]

Figure 1 shows the architecture of an LSTM Network. The LSTM is comprised of four main gates, an input gate, "a forget gate", an update gate, and finally an output gate. The gates essentially determine which data the model should remember and which data isn't useful for training and can be forgotten. The models trained achieved a much better result than the model in the previous study and when considering the much larger size of the dataset the model trained as part of this study would have a much higher accuracy when it came to generalizing and classifying new data.

In the literature review we also address the challenges and limitations of using artificial intelligence in developing automated diagnosis systems for COVID-19 classification. In a paper by Huang, Yang et al [4], researchers discuss the current challenges and limitations of developing an artificial intelligence model to assist medical professionals in the identification and diagnosis of COVID-19. The challenges include: lack of data, lack of data quality, and the use of poorly merged datasets which we discussed earlier ("Frankenstein datasets"). There are more challenges other than those previously discussed such as finding patients who are COVID-19 positive and asymptomatic (most will not seek medical diagnosis as they don't present symptoms). The labelling of data is also an issue as there is a risk of mislabelled data by clinicians given moderate signs of COVID-19 being hard to definitively classify as COVID positive or negative. The risk of false positives and false negatives also present issues such as unnecessary quarantine and further spread of infection. To mitigate these challenges Huang, Yang et al suggest that when developing an AI diagnostic system that developers should combine chest imaging, exposure history, and laboratory tests when training / testing the model. The issue with this approach is laws concerning data collection and ethical questions regarding the right to the patients privacy. There is also a high cost associated with data collection and collecting a sufficient amount of data may take quite a long

time.

The issues explored in the literature review discuss the use of CNNs and GANs across different problem domains. The uses and limitations of AI are also discussed in this section as are the tactics and techniques researchers have used to improve model performance. In the literature review section we research and explore the use of GANs in the creation of synthetic data in a number of different areas and to examine the current paradigms in utilizing GANs to create synthetic examples from the datasets. We then discuss the key advantages of using GANs: larger datasets to train models, eliminates the cost of data collection and labelling, and many more advantages. In addition we also discuss the disadvantages of GANs such as the lack of diversity in data created and an inability to reduce bias in the dataset. We also examined a number of CNN architectures and their particular advantages and disadvantages when utilized in lychee detection and COVID-19 diagnosis.

After discussing the advantages and limitations of using synthetically augmented data we then moved on to discuss a paper by Tanaka and Aranha[5]. In this paper the researchers discuss two algorithms SMOTE (Synthetic Minority Over-sampling Technique) [6]and ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning)[7]. Both techniques are used to oversample the minority class within datasets to balance the dataset.

We then discuss how each algorithm works. SMOTE works by creating artificial data which takes into account the data's position, a random point in the least represented class in the dataset is selected and SMOTE identifies members of the same class within the data by using the k-nearest neighbour algorithm which is a form of unsupervised learning. SMOTE then generates an entirely new point in the vector for each pair which is situated between the two pieces of data, the new point is then positioned at a random percentage away from the initial point chosen[5]. ADASYN works in a similar way to SMOTE and was originally based on SMOTE. Both function in much the same way but the key difference lies in ADASYN adding a random small bias value to the points, breaking linear correlation to their parents. The bias that ADASYN adds helps to increase the amount of variance within the synthetic data.

We then evaluate the performance of each of these algorithms. When testing both SMOTE and ADASYN the researchers found that both underperformed in accuracy on the imbalanced dataset compared to the original model but had a higher recall. When the datasets were balanced it was found that SMOTE and ADASYN performed much better. The models also achieved an increase in accuracy and performance when trained using traditional GANs. By reviewing this study it showed that the use of GANs were able to improve accuracy of certain classification models across a number of problem domains. The researchers used 3 datasets when training the models and synthetically augmenting the datasets. These datasets were

- Pima Indians Diabetes data Database
- Breast Cancer Wisconsin Dataset (Diagnostic)
- Credit Card Fraud Detection

Given that the researchers had shown success in using GANs across these vastly different problem domains it showed that there may be some value into researching syntehtic data augmentation when it came to alleviating data shortages in COVID-19 diagnosis.

The next study analyzed as part of researching the current use of GANs to augment imbalanced datasets was from a paper by Wang and Xiao[8]. In this paper the use of convolutional neural networks was employed to discern defective lychee. The dataset used was augmented using a generative adversarial network to solve the issues of imbalance between classes in the dataset used. The researchers also used a variety of CNN architectures which included: SSD-MobileNet V2, Faster RCNN-ResNet50, and Faster RCNN-Inception-ResNet V2. The models used were trained with different hyperparameters to evaluate their performance. Through the use of GANs the researchers were able to increase the mean average precision for each of the following architectures: SSD-MobileNet V2 2.86%Faster RCNN-ResNet50 V2 1% Faster RCNN-Inception-ResNet V2 0.58%. The mean average precision gap was also greatly reduced through the use an augmented dataset which can be seen in Table I.

Name of Model	Mean Average Precision Performance Gap
SSD-MobileNet V2	1.78%
Faster RCNN-ResNet50 V2	4.45%
Faster RCNN-Inception-ResNet V2	2.35%

TABLE I: Mean average precision after Augmentation(lychee Surface Defect Detection Based on Deep Convolutional Neural Networks with GAN-Based Data Augmentation)[8]

When evaluating the performance of the models the researchers found that GAN augmentation had increased the accuracy of a number of models as shown in Table II.

Model	Setting	Acc	Rec	Spe	F1
SSD-MobileNet V2	Base setting	89.81%	90.08%	89.89%	89.46%
SSD-MobileNet V2	GAN Augmentation	91.96%	92.06%	91.99%	91.92%
Faster RCNN-ResNet50	Base Setting	91.82%	92.23%	91.95%	91.72%
Faster RCNN-ResNet50	GAN Augmentation	92.76%	92.96%	92.80%	92.55%
Faster RCNN-Inception-ResNet V2	Base Setting	91.96%	92.07%	91.98%	91.54%
Faster RCNN-Inception-ResNet V2	GAN Augmentation	92.36%	91.74%	92.22%	91.86%

TABLE II: Comparison of accuracy of base models vs models with data augmentation(lychee Surface Defect Detection Based on Deep Convolutional Neural Networks with GAN-Based Data Augmentation)[8]

From the studies analyzed in the literature review it is clear to see that data augmentation has achieved promising results in certain problem domains. Through the literature review we gained a perspective and insight into how to go about balancing imbalanced datasets when it came to augmenting COVID-19 X-rays and CT Scans. The literature review revealed key issues with current automated diagnostic models for COVID-19 and also showed that through the use of GANs we could solve some of the issues affecting current models.

IV. Design & Implementation

When starting off the research we decided it would be best to include baseline CNN models which we would use as a

metric when evaluating the performance of the augmented models. The baseline models used a train / validation / test split of 70% for training, 10% for validation and 20% for test. The reasoning for this is that a large amount of data should be used for training as the model would need to be exposed to many features in the data. The validation split was important to determine if the model was performing well during training. The test set size was double the validation size so that when gauging the model's performance the test set would have more variety and could better measure the model's ability to generalize. The reasoning for the use of baseline models was so that we could have a metric when it came to comparing the baseline models and baseline pre-trained models with the synthetic models. The use of transfer learning was also employed when implementing the CNNs. Three additional models were created and trained on each dataset we used the following CNN architectures: Xception, ResnetModel50V2, and EfficientNetV2S. Each of these architectures were trained on the ImageNet dataset, which is composed of 14,197,122 images and contains 1,000 classes [9]. Each of these pre-trained models also had an additional 2 layers appended to them, one layer to train and another layer which was the output layer used for classification.

After achieving satisfactory results on these baseline models, we then set to work upon researching and creating the GAN models. The DCGAN models were created based off of an example from the Keras website [10]. The models were run a number of times with hyperparameters being adjusted to improve the model's performance. After examining the images and comparing and contrasting them to real samples selected from the dataset, it was clear that the models were in fact producing a similar output to the real samples in the dataset.

Fig. 2: Real COVID-19 Radiography Mask Example

Fig. 3: Generated COVID-19 Radiography Mask Example DCGAN

As shown in the above Figures 2 and 3 the real and synthetic example share a striking resemblance. There were however a few issues with the results produced from the GANs. There were a number of malformed images, some which had artefacts which can be seen in figure 4 and figure 5

Fig. 4: Synthetically generated COVID 19 mask with Artefacts(DCGAN)

Fig. 5: Synthetically generated COVID 19 mask with Artefacts(DCGAN)

The next Figures show the synthetically generated X-ray example 7 lacks the quality of the original 6 but appears to have similar features. The next figures show an example of a synthetically generated pneumonia X-ray along with a sample taken from the dataset

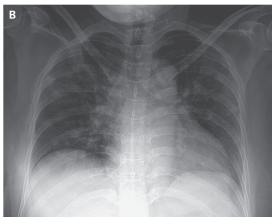

Fig. 6: Real Pneumonia X-ray Dataset COVID-19

Fig. 7: Synthetically Generated Pneumonia X-ray Dataset COVID-19 DCGAN

These issues occurred across the datasets and affected all the images produced by the GAN, regardless of if the images were X-rays or CT scans. Some of the images produced by the GANs also appeared slightly grainy or lacking in quality compared to those in the dataset. This issue may have been caused due to the downscaling in resolution as due to computational limitations the images produced by the GANs in this study have a resolution of 128×128. The amount of malformed images may have been greatly reduced by increasing the resolution of the images output by the GANs and standardising the images input into the GAN. It should be noted a number of datasets used in this study did not have a standardised resolution for images.

The use of VAEs was also employed for this study but they did not seem to produce satisfactory results. Most of the images produced by the VAEs appeared to be the exact same image, which leads us to believe the VAE suffered from mode collapse.

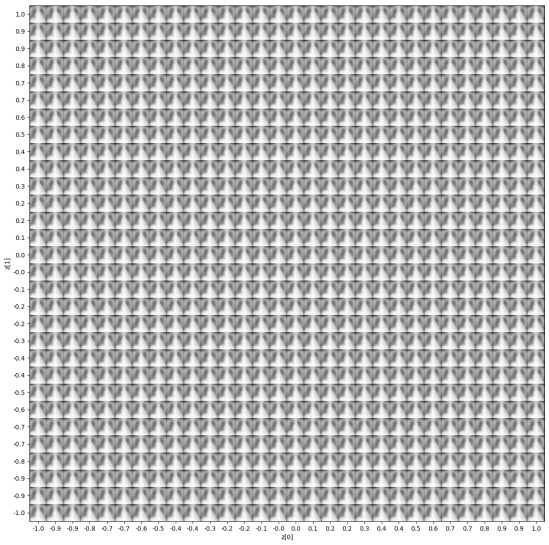

Fig. 8: Example of VAE mode collapse

After training multiple VAE models it appeared that some suffered from mode collapse and others produced no output at all. There may have been some use to the images produced by the VAE but to the untrained eye each image produced looks the same.

Once the GANs were trained the discriminator and generator models were then loaded and were run for as many epochs as it would take to produce the number of images required. Due to computational limits and the need to make the imbalance between classes as small as possible, the GAN produced 10 images per epoch and brought the classes into relative balance with each other.

The images were then saved and were used to augment the different datasets by moving the produced images into the target dataset. Once this was done the original baseline CNNs were replicated and the data reloaded and split so that our new data would be used in the training,validation, and testing of the models. The train / validation / test split was kept the same, with 70% of data used for training, 10% used for validaiton and 20% used for the test set. The split was kept the same so the results would be objective. The synthetic data was present in both the validation and synthetic data for datasets which were of a sufficient size. The only dataset which did not contain any synthetic examples in the training or validation dataset was the X-ray COVID-19 dataset.

Once the new models were trained and their performance documented we then moved on to comparing and contrasting the results with those of the baseline models. The new models were exact copies of the previous baseline models, the only difference being that they were trained using the augmented datasets.

V. Results

Overall the results from this research showed that satisfactory synthetic data can be produced by GANs to help augment and balance existing datasets in this domain. It is evident

when viewing the accuracy of the augmented models that they were able to achieve better performance and higher train / validation / test accuracy scores. There were, however, a number of models which performed slightly worse than the existing models. There could be many reasons for this such as artefacts in the synthetic images, malformed synthetic images, and also the increase in the dataset size. The top models which performed better than the originals when augmented were the Extensive CNN CT EfficientNetV2S model which achieved a test score accuracy of approximately 98% which had an increase of approximately 5% when compared to the original model which has a test set accuracy of approximately 94%. The Radiography EfficentNetV2S Model also improved greatly in terms of accuracy and had a final test set accuracy of approximately 96% which is an improvement of approximately 8% when compared with the previous accuracy of 88% which was attained by the original model.

We have decided to include the plots for the accuracy and losses of the top models below so that they can be compared and contrasted with each other. Figure 9 shows the non augmented training and validation accuracy and figure 10 shows the training and validation loss for the Extensive COVID 19 Dataset CT models.

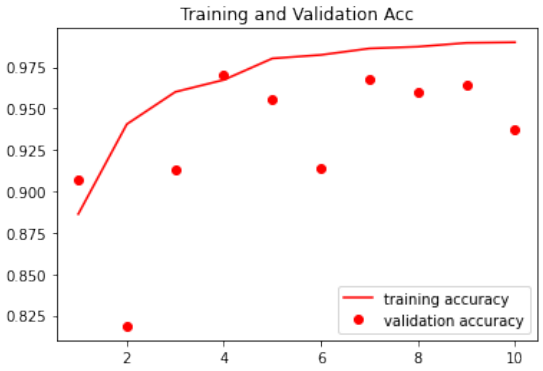

Fig. 9: Transfer Learning EfficientNetV2S CNN Baseline Train and Validation Accuracy Extensive COVID 19 Dataset CT - The X-Axis shows the epoch number and the Y-axis shows the accuracy

Fig. 10: Transfer Learning EfficientNetV2S CNN Baseline Train and Validation Loss Extensive COVID 19 Dataset CT - The X-Axis shows the epoch number and the Y-axis shows the loss

Figure 11 shows the augmented training and validation accuracy and figure 10 shows the training and validation loss for the Extensive COVID 19 Dataset CT.

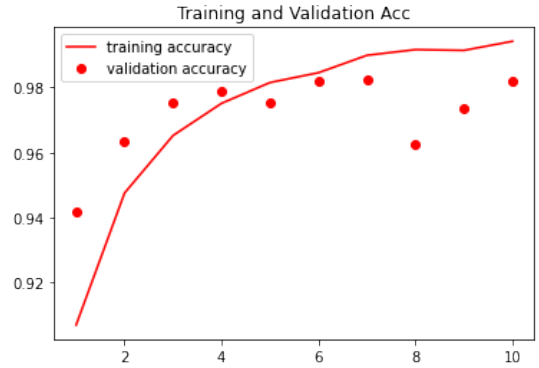

Fig. 11: Extensive CT Augmented EfficientNetV2S Model DCGAN Accuracy - The X-Axis shows the epoch number and the Y-axis shows the accuracy

Fig. 12: Extensive CT Augmented EfficientNetV2S Model DCGAN - The X-Axis shows the epoch number and the Y-axis shows the loss

Figure 13 shows the non augmented training and validation accuracy and figure 14 shows the training and validation loss for the Radiography models.

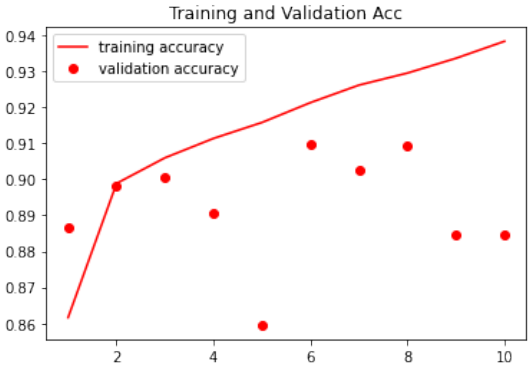

Fig. 13: Transfer Learning EfficientNetV2S CNN Baseline Train and Validation Accuracy - The X-Axis shows the epoch number and the Y-axis shows the accuracy

Fig. 14: Transfer Learning EfficientNetV2S CNN Baseline Train and Validation Loss - The X-Axis shows the epoch number and the Y-axis shows the loss

Figure 15 shows the augmented training and validation accuracy and figure 16 shows the training and validation loss for the Radiography models.

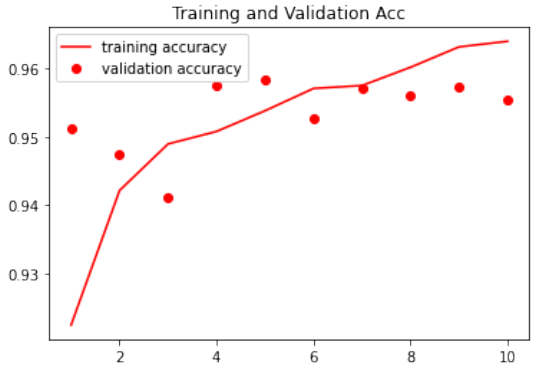

Fig. 15: Radiography Augmented EfficientNetV2S Model DC-GAN Accuracy The X-Axis shows the epoch number and the Y-axis shows the accuracy

Fig. 16: Radiography Augmented EfficientNetV2S Model DC-GAN Loss - The X-Axis shows the epoch number and the Y-axis shows the loss

From viewing the above plots it is shown that the augmented models had a decrease in loss and an increase in training accuracy.

The improvements in accuracy are not the only factor to take into account when evaluating these models. The more data the models are exposed to the more features they can be trained to detect when diagnosing a patient. Given that the datasets were augmented by images which number in the thousands, it can be assumed that the models trained in this study are more robust and able to generalize and adapt to new data better than those trained on the original datasets.

Judging the synthetically generated output from the GAN it was shown that the GAN produced some malformed images and images with artefacts which can be seen in Figures 5 and 4. However, the GANs also produced some convincing images which share a lot of the same features as the examples taken from the dataset as we can see from Figures 3 and 7. Overall the GAN models trained produced images which were remarkably similar to the images within the datasets from which it was trained.

VI. Conclusion

In conclusion, despite the computational resource and the limitations faced throughout the course of this study, a number of models showed significant improvement in terms of both accuracy and loss when the data was augmented. Although none of the original or augmented models outperformed the accuracy seen in the models evaluated in the literature review, it is worth noting that the models did achieve a high test set accuracy, the models in the literature review were not evaluated using a test set and may have overfitted to the validation set. The results of this study are significant in that they show that through the use of synthetic data-augmentation the overall performance of models could be improved in this problem domain. The model's ability to generalize is very important in the diagnosis of COVID-19 as it can help medical professionals prevent the spread of COVID-19 and would help reduce the risk of unnecessarily quarantining patients who were incorrectly diagnosed as COVID positive.

The conclusion of this research shows that there is much more work to be done in this problem domain. To truly evaluate the models included in this study they would have to be deployed and evaluated in a clinical setting so their usefulness can be assessed by medical professionals. There are numerous improvements which could be made to this study such as those mentioned earlier in this paper, such as data pruning and evaluation of the GAN output. The overall objective of the research was achieved and the usefulness of synthetic data has been proven by evaluating the results of the baseline models with the results of the augmented models.

Although there are many more improvements which could be made to this study the results show that the balancing of classes through the use of generative deep learning can produce propitious results and more research into this area is needed.

References

[1] I. Islam and Asraf, "A Combined Deep CNN-LSTM Network for the Detection of Novel Coronavirus (COVID-19) using X-ray Images," *Informatics in Medicine Unlocked*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820305621.

[2] B. Mahmoudi *et al.*, "A Deep Learning-Based Diagnosis System for COVID-19 Detection and Pneumonia Screening Using CT Imaging," *Applied Sciences*, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/10/4825/.

[3] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 06, no. 02, pp. 107–116, 1998. [Online]. Available: https://doi.org/10.1142/S0218488598000094.

[4] Y. Huang *et al.*, "Artificial Intelligence in The Diagnosis of COVID-19 Challenges and Perspectives," *International Journal of Biological Sciences*, pp. 1–7, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33907522/.

[5] Tanaka and Aranha, "Data Augmentation Using GANs," *Proceedings of Machine Learning Research*, pp. 1–16, 2019. [Online]. Available: https://arxiv.org/pdf/1904.09135.pdf/.

[6] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," *CoRR*, vol. abs/1106.1813, 2011. arXiv: 1106.1813. [Online]. Available: http://arxiv.org/abs/1106.1813.

[7] G. He Bai, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *IEEE World Congress on Computational Intelligence*, pp. 1322–1328, 2008. [Online]. Available: https://www.researchgate.net/publication/224330873_ADASYN_Adaptive_Synthetic_Sampling_Approach_for_Imbalanced_Learning.

[8] Wang and Xiao, "Lychee Surface Defect Detection Based on Deep Convolutional Neural Networks with GAN-Based Data Augmentation," *Agronomy*, vol. 11, 2021. [Online]. Available: https://www.mdpi.com/2073-4395/11/8/1500.

[9] Multiple, "ImageNet," [Online]. Available: https://www.image-net.org/index.php/.

[10] F. Chollet, "DCGAN to Generate Face Images," [Online]. Available: https://keras.io/examples/generative/dcgan_overriding_train_step//.