Automatic Detection of COVID-19 Pneumonia in Chest Computed Tomography Scans Using Convolutional Neural Networks

Neil Micallef
Artificial Intelligence Dept.
University of Malta
Msida, Malta
neil.micallef.14@um.edu.mt

Carl James Debono
Communications and Computer Engineering Dept.
University of Malta
Msida, Malta
c.debono@ieee.org

Dylan Seychell
Artificial Intelligence Dept.
University of Malta
Msida, Malta
dylan.seychell@um.edu.mt

Conrad Attard
Computer Information System Dept.
University of Malta
Msida, Malta
conrad.attard@um.edu.mt

2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON) | 978-1-6654-4280-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MELECON53508.2022.9843100

***Abstract*—The Coronavirus outbreak caused by the SARS-CoV-2 virus has been the focal point of global attention over the past two years, owing to the pandemic's infection rate and the huge burden on the world's healthcare systems and economy. Diagnosis of infection by the virus may be carried out through a number of tests, with the current mainly used technique being reverse transcription polymerase chain reaction tests. An alternative approach for diagnosis is through the use of medical imagery such as chest X-rays or chest Computed Tomography images. In this work, we propose a machine learning driven approach which automatically detects pulmonary pathological features caused by the Coronavirus infection in chest Computed Tomography images. The model was trained and evaluated on the COVIDx CT-2A dataset, achieving an accuracy of 96.31% on the testing segment of the dataset.**

***Index Terms*—COVID-19, Coronavirus, computed tomography, deep learning**

I. Introduction

The Coronavirus disease can be diagnosed using reverse transcription polymerase chain reaction (RT-PCR) tests which require swabs to be collected from individuals, including patients without symptoms. These tests detect nucleic acid from the virus in the respiratory specimens, flagging the presence of the virus. The detection accuracy of this method depends on the load of virus in the specimen and hence there are instances where the virus, although present, remains undetected. Real-time RT-PCR test results have been shown to have inconsistencies when assessed on their own. This is shown in work such as [1], where it is proposed that real-time RT-PCR test results should be used in conjunction with clinical features, especially Computed Tomography (CT). CT images are also among the most researched techniques when combining AI with radiology [2]. In the context of this work, computer vision algorithms can be applied on the CT imagery to facilitate and support the diagnosis process, with Convolutional Neural Networks (CNN) being a standard tool. Thus, to implement this we present a simple CNN model that is capable of obtaining comparable results with the state-of-the-art when distinguishing between CT images with COVID-19 pneumonia, common pneumonia, and normal controls. Moreover, the model is required to minimize the number of false positives and false negatives to provide an accurate diagnosis for each patient.

Relevant work in the domain is outlined in Section II below. Following a description of the related literature, Section III provides a detailed breakdown of the techniques used, including the CT image dataset, CNN architecture, and how model training was conducted. The results obtained by the model in our experiments are then discussed in Section IV and compared with the current state-of-the-art for this dataset. Section V presents opportunities for potential enhancements to this work and the conclusions derived from this study.

II. Background

Although the COVID-19 outbreak is a recent phenomenon, there are already a number of works presenting Machine Learning (ML) and Deep Learning (DL) approaches to detect COVID-19 from medical imagery. A number of review papers have been presented which show the strengths and weaknesses of these approaches at a holistic level. The work in [5] emphasises the importance of dataset selection and having additional validation datasets, also echoed in [8]. Other limitations are discussed in the two-part work in [6], [7], such as biases introduced in datasets from having individuals within the same age or sex groups. Important markers for classification may also be removed when applying preprocessing techniques and data augmentation on image datasets. Shortcut learning is also discussed in detail in the second part of the work in [7], where examples are mentioned such as models making use

978-1-6654-4280-0/22/$31.00 ©2022 IEEE

of features outside of the main subject as a basis for their classification. [8] mentions the disadvantage of deep neural networks requiring vast amounts of data to learn correctly. However, they also argue that deep CNNs provide the most accurate technique for COVID-19 classification, stating that more such works need to be explored.

Examples of implementations of deep CNNs include the work by Panwar *et al.* [3], who built a CNN model with the popular VGG16 architecture to detect the presence of COVID-19 from chest X-ray images. The authors report accuracies of ~97% with a sensitivity of 97.62% and specificity of 78.57% [3]. The work in [4] presented a CNN model whose input consists of CT images, obtained positive results and also identified 17 of 25 positive patients who had been diagnosed as negative by radiologists. A drawback of the work in [4] is that the model is trained on data solely from medical centers in Chinese provinces.

The COVIDx-CT dataset [9] used in this study contributed to models which achieve state-of-the-art performance for the detection of COVID-19 pneumonia. The dataset authors first presented their CNN model COVIDNet CT-1 [10], pre-trained on the ImageNet dataset and reported a 99.1% accuracy on the original COVIDx-CT data. It is noteworthy that an extensive array of data augmentation techniques were applied to the original dataset. They later proposed a new CNN named COVIDNet CT-2 [11], assessing its performance on the latest iteration of the COVIDx-CT dataset, COVIDx-CT 2A/2B. COVIDNet CT-1 [10] accuracy dropped to 94.5% in this latter study, and was surpassed by COVIDNet CT-2's best score of 98.1%.

These works [10], [11] are notable not only for their positive scores on a test dataset of substantial size, but also for the models being small in terms of the number of training parameters. There are some considerations to these works. Whilst it is a best practice for AI practitioners to supplement datasets with limited samples, for this particular problem it may result in the removal of important classification markers as discussed in [7]. Moreover, whilst pre-training on ImageNet appears to have contributed positively to the model's scores, the weights and activations learned from features outside of the medical domain could result in bias. Nonetheless, [11] show that their approach makes predictions based on radiological features via explainability.

III. Methodology

A. Data Definition

In problems related to medical imaging, scarcity of data and unbalanced datasets are often one of the major struggles linked to the research. In spite of this and also the fact that the COVID-19 pandemic began very recently, there are a number of well-defined datasets with a comprehensive number of images and ground truths. As mentioned in Section II, we made use of the open-source dataset found in an online Kaggle competition [10], [11], with the data available for download on the webpage of the challenge [9]. We used the COVIDx CT-2A version of the dataset, as the ground truths in this version are all well verified using RT-PCR test results or radiologist diagnosis. The COVIDx CT-2A dataset features over 194,000 CT Scans from 3745 patients, with classification split into three labels, with label 0 corresponding to healthy images, label 1 to common pneumonia, and label 2 to COVID-19 pneumonia. The proportion of images under each label are shown in Table I.

TABLE I
Distribution of classes for the training, validation, and testing sets of the COVIDx CT-2A Dataset. The training set contains a substantial amount of COVID-19 cases, while the validation and testing data contain more realistic evaluation grounds with a larger proportion of healthy samples to mimic actual scenarios.

Dataset Split	Healthy (Label 0)	Pneumonia (Label 1)	COVID-19 Pneumonia (Label 2)	Total
Train	35,996	25,496	82,286	143,778
Validation	11,842	7,400	6,244	25,486
Test	12,245	7,395	6,018	25,658

From Table I, one may observe how the training dataset has a bias towards COVID-19 samples. Whilst this could be beneficial to increase the model's prediction accuracy for the COVID-19 class, we opted for a resampling strategy which resized each of the classes in the training and validation data to the class with minimum representation, which was label 1 (common pneumonia). The motivation behind this strategy was to perform model training without lending bias to any particular class. The testing dataset was the only set of images that was not resampled in this way, in order for the evaluation to be conducted on the full test set. Keeping the test dataset in its entirely also assists in assessing how the model handles the imbalance in the test dataset which favours healthy samples, as this reflects the real-life scenario for this problem.

B. Model Architecture

The model proposed here and shown in Fig. 1 is adapted from a 3D CNN architecture that was used to predict viral pneumonia in chest CT Scans [12]. This model was chosen to have a solid and workable basis for adaptation. Furthermore, the final model presented here is still straightforward enough for other researchers to emulate and adapt even further. There are a number of differences in our adaptation's architecture and training parameters compared to [12], starting with the input modality, as the input to our solution is supplied as 2D images. Moreover, we double the filter resolution to range from [64..1024] rather than [32...512]. The original approach was focused on a dataset which only presents 'normal' and 'abnormal' CT scans, whilst our classification is multiclass across the healthy, pneumonia, and COVID-19 pneumonia target classes. As a result, we use a softmax activation function for the final classification in place of the original's sigmoid and also use the sparse categorical crossentropy as a loss function in line with our input class structure.

The CNN presented in this work makes use of 2D convolutions with 3×3 kernels and standard ReLU activations,

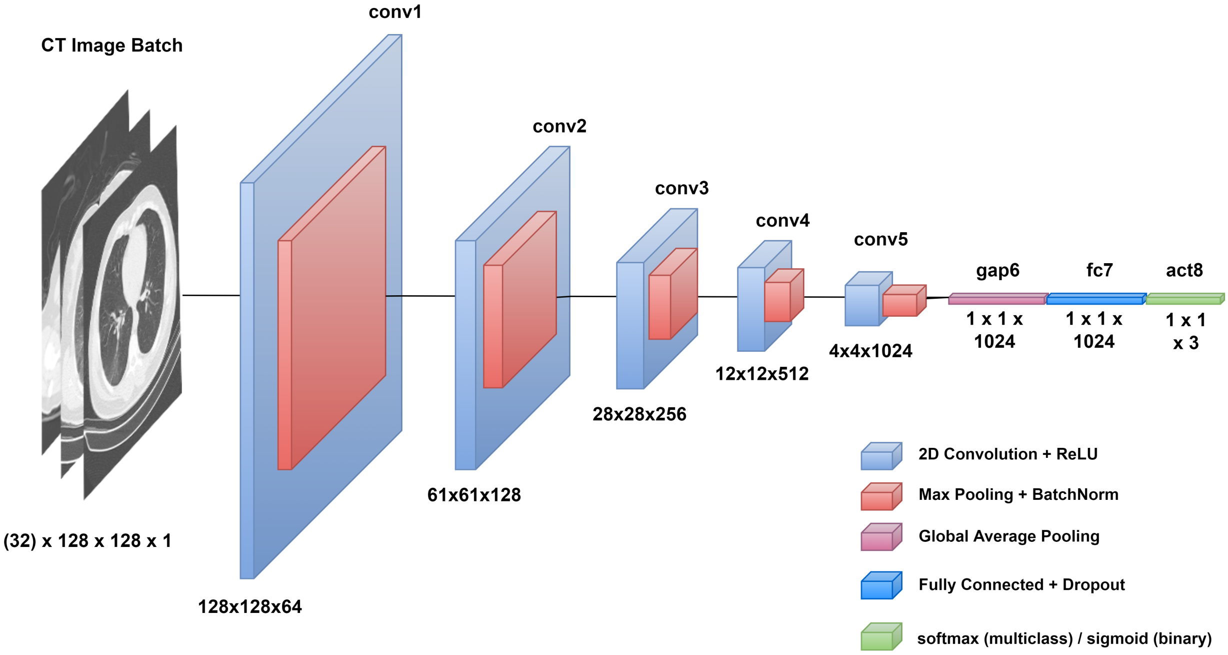

Fig. 1. Diagram showing the model architecture for the work presented in this study. The input resolutions for each block are shown under each section of the diagram. The global average pooling layer flattens the feature information in preparation for the fully connected layer. The activation function is either softmax or sigmoid depending on whether the training cycle is multiclass or binary.

max pooling, and batch normalization layers. The output at the maximum depth of the network is then flattened into a single-dimensional vector and passed through a softmax activation function to output the multiclass prediction probabilities. A single dropout layer is applied only to the one-dimensional flattened output as per the original network. We did not include dropout for the 2D convolutional layers, observing Hinton *et al.*'s [13] claim that dropout appeared to produce less substantial results in CNN convolutional layers. The depth and filter resolution of the final network resulted in a total of 7.3M model parameters.

C. Model Training

The data resampling process discussed in Section III-A is the first step in the pipeline. Keras image generators are then populated with the data, applying 20 degree rotations to the training image batches to increase the variety of input samples. Following data loading, the images were passed to the convolutional neural network (CNN) model used for this experiment. Additional preprocessing was not applied to the images, following the starter Kaggle notebooks and work by the authors of [10].

The Adam Gradient Descent optimizer [14] was used, keeping the default parameters barring the learning rate α. The initial α value tested for training was $1e^{-4}$, however this was reduced to $1e^{-5}$ to stabilize the learning curve. Subsequent training runs showed that model training converges between 13 and 20 epochs. Thus, 20 epochs was set as the maximum, and an early stopping strategy was implemented with a patience of 5 to stop training past convergence. The training cycle for the best performing model following the process above is shown in Fig. 2. The learning curve shows that training was stopped at epoch 19 by the early stopping mechanism to counter overfitting, since the highest validation accuracy was obtained in epoch 14.

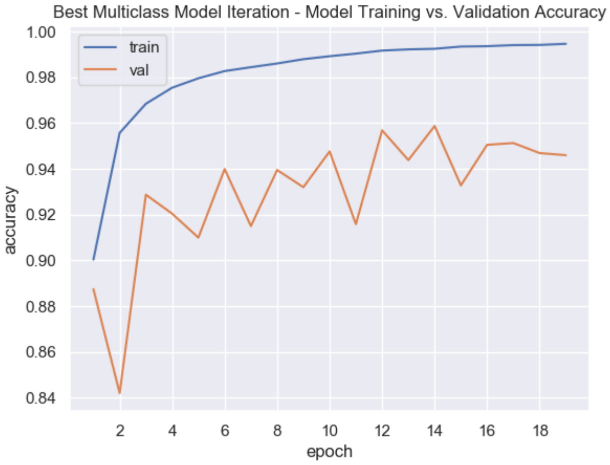

Fig. 2. Model training cycle for the best model iteration as discussed in Section IV-B below.

The loss function used for multiclass classification was sparse categorical cross-entropy, since the target classes are in the set $y \in \{0, 1, 2\}$. The sparse categorical classes of 0, 1, and 2 meant that COVID-19 was being given the highest 'rank' (as a class) by the model during training. This was maintained in favour of a 'one-hot' encoding categorical approach to still give slight priority to the 'COVID-19' class as a target. Training was performed on a MSI GS60 6QE laptop with 32GB of RAM, a 7^{th} generation CPU, and a GTX970 GPU. Since the model was trained on over 76,000 training images following resampling, this resulted in a training time of ~15-20 minutes per epoch. The model was trained five separate times on the same seed of randomly generated image batches to ensure that the final results were consistent across separate training runs.

It is also noteworthy that bounding box coordinates were provided for each sample as part of the dataset, aiming to facilitate background removal for each image by isolating the region of interest in a CT scan. Since the bounding boxes follow the contours of each patient's scan, their dimensions are not uniform. Moreover, cropping some samples to these dimensions can result in skewed aspect ratios. CNNs also generally work better with square input images. Since the bounding boxes were provided, we still attempted a multiclass training cycle using images cropped to the given coordinates. However, the evaluation was obtaining random results unless done on an uncropped version of the test dataset, where the model still obtained worse results compared to those in Section IV below. As a result of all the above factors, the bounding box approach was not investigated further.

IV. Results

This section first discusses the evaluation metrics used in this study. It then presents the results obtained for the multiclass and binary class solutions, and an analysis of the multiclass misclassifications. This is followed by an evaluation of these results compared with related research.

A. Evaluation Criteria

The evaluation criteria used for this study support accuracy as a metric by also providing an emphasis on sensitivity and specificity. These two metrics are essential to understand the extent of this system when applied in a real-world context. Accuracy is a standard metric used for model evaluations across many domains, thus providing a means to assess performance against other research. The second metric used for the evaluation is the f1 score. The f1 score is more useful for imbalanced class distributions since its formulation emphasizes negatives more than accuracy. We utilize a weighted version of this metric which calculates a weight based on the true-valued cases of each class. Nonetheless, both accuracy and f1 score may present an overoptimistic score depending on the data distribution, as shown in [15]. We conducted the evaluation with this in mind, also employing a multiclass version of the Matthews Correlation Coefficient (MCC):

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}} \tag{1}$$

where c refers to the number of correct predictions, s refers to the number of samples, t_k refers to the number of actual occurrences of class k, and p_k refers to number of predictions of class k. The combination of accuracy, f1 score, and MCC assesses the model both in terms of how well the model generates predictions, and also accounting for minimization of false diagnoses through the f1 score and MCC.

B. Classification Results

The predictions generated by the model for the test dataset were recorded for five separate instances of the model architecture discussed in Section III-C, and compared against the ground truths. The class probabilities generated by each model were also averaged into a single prediction vector, rounded to the nearest class, and evaluated in a separate comparison. The results are shown in Table II.

TABLE II

THE EVALUATION RESULTS OF THE MULTICLASS MODEL AGAINST THE GROUND TRUTHS OF THE 25,658 IMAGES IN THE COVIDX CT-2A TEST DATASET. THE BEST SCORES ARE MARKED IN BOLD FONT.

Instance	Epochs	Accuracy (Train)	Accuracy (Validation)	Accuracy (Test)	F1 Score	MCC
1	15	0.9894	0.9459	0.9429	0.9434	0.9120
2	18	0.9928	0.9543	0.9567	0.9569	0.9322
3	8	0.9680	0.9387	0.9388	0.9397	0.9065
4	19	0.9925	0.9588	**0.9631**	**0.9630**	**0.9419**
5	20	**0.9934**	**0.9603**	0.9577	0.9577	0.9334
Averaged	-	-	-	0.9486	0.9487	0.9205

The results reported in Table II show that the model's scores are very consistent across separate training runs. The model with the largest deviation was the third instance, where training was terminated early due to the early stopping callback's patience parameter being exceeded early on. The scores for the best instance, which is the fourth, exhibit positive scores throughout, with only a slight drop from the f1 score and accuracy to the MCC, which still reported a score of 94.19% calculated on the entire test dataset. In order to understand the strengths and weaknesses of the best instance, the confusion matrix is shown in Fig. 3. The column values of the confusion matrix are investigated further in Section IV-C below.

We also conducted a binary classification experiment, omitting the common pneumonia label from the problem. The main outcomes of this experiment were that both the normal controls and the COVID-19 samples had a slight increase in prediction scores. The best evaluation scores were obtained by the averaged predictions of five trained models on the binary dataset, with an improved test accuracy of 97.09%. Whilst the improved result is a definite positive, it is possible that the increase in class-specific training samples for the remaining two classes could have been a contributing factor to the binary model's improved performance.

C. Analysis of Misclassifications

The top misclassifications of the best multiclass model iteration (shown in Fig. 3) can be split into three groups: healthy images predicted as COVID-19 pneumonia images (1.06%); COVID-19 pneumonia images predicted as healthy images (0.96%), and COVID-19 pneumonia predicted as common

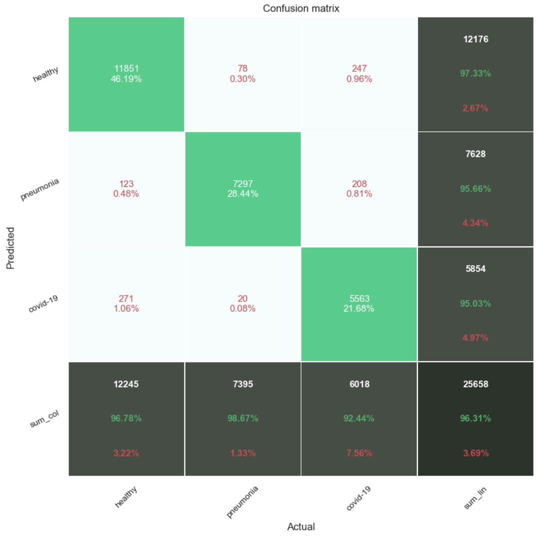

Fig. 3. Confusion matrix for the prediction results on the separate test dataset using the best multiclass model instance as reported in Table II. The white numbers on green and red numbers on white represent correctly and incorrectly classified images. The percentage values accompanying each number represent the value's percentage from the entire test dataset.

pneumonia (0.81%). From these scores, an experiment was conducted to check which data sources in the test dataset contributed to the incorrect predictions, shown in Table III.

TABLE III
ANALYSIS OF INCORRECT PREDICTION GROUPS ON THE TEST DATASET. THE NAMING CONVENTION FOR COLUMNS 3 TO THE PENULTIMATE COLUMN IS "PREDICTED CLASS LABEL - GROUND TRUTH CLASS LABEL". THE FINAL COLUMN SHOWS THE TOTAL PERCENTAGE OF INCORRECT PREDICTIONS FOR THE CORRESPONDING DATA SOURCE.

Source	Test Cases	0-2	2-0	1-2	0-1	1-0	2-1	% Incorrect
CNCB	21190	32	13	100	102	78	20	1.6281
COVID-CTSet	2411	212	152	72	2	0	0	18.1667
COVID-19-20 /TCIA	887	0	51	36	0	0	0	9.8083
LIDC-IDRI	841	27	0	0	19	0	0	5.4697
coronacases.org	254	0	17	0	0	0	0	6.6929
radiopaedia.org	75	0	14	0	0	0	0	18.6667

The results presented in Table III show that although CNCB accounted for more than 82% of the test dataset, the number of false predictions was lower than that of the COVID-CTSet. This is likely since CNCB also formed a large segment of the training data, leading to the model having a higher generalization for its images. The other data sources' results are as expected, since the datasets have a smaller number of patients in the training set. Moving back to the COVID-CTSet's results, these are interesting as COVID-CTSet was actually the third largest repository in our resampled training split. The main misclassification groups for this data are 0-2 and 2-0, meaning COVID-19 false positives and negatives.

To gain further insight into the predictions, we made use of activation visualization techniques. Our main considerations were (Grad-CAM) [16] and guided saliency maps [17]. Whilst Grad-CAM is more descriptive in general, the final convolutinal layer in our model is a number of layers' distance from the

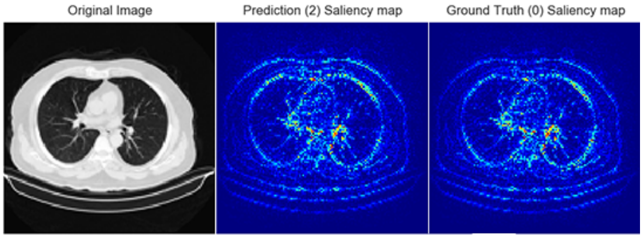

Fig. 4. Guided saliency map for a patient from the COVID-CTSet data with normal controls predicted as having COVID-19 pneumonia.

final class prediction layer, which results in Grad-CAM being less accurate. On the other hand, saliency maps are applicable to Dense layers, such as our model's final classification layer. An example of a guided saliency map on an affected sample from the COVID-CTSet is shown in Fig. 4.

The maps were generated using guided backpropagation saliency maps as presented in [18]. From the results gathered on the other images with false predictions, there appears to be a pattern of edges being the main discriminative feature. Another observation we made was that elements from outside the body (such as the machine artifacts in the dataset) were not being highlighted by the saliency maps. Since there are no visual abnormalities in the image, it could be that the high intensity areas near the center of the image were incorrectly interpreted by the model as abnormalities.

D. Evaluation

TABLE IV
COMPARISON OF MODEL PREDICTION ACCURACY ON THE TESTING SPLIT OF THE COVIDX-CT 2A DATASET.

Network	Accuracy
Ours	0.963
COVID-Net CT-1 [10]	0.945
COVID-Net CT-2S [11]	0.979
COVID-Net CT-2L [11]	0.981
Zhao *et al.* [19]	0.992

Whilst there are a number of studies presenting COVID-19 image classifiers, each dataset in the domain has very specific nuances and may also aim to solve different problems. Given the size and scope of the COVIDx CT-2A dataset, the best comparisons would be drawn against COVID-NET CT-1 [10] and COVID-NET CT-2 [11], reported by the providers of the dataset. The metric which is common to both these works and ours is the accuracy when evaluating the model on the test dataset. This comparison is shown in Table IV.

It is also noteworthy that since this dataset was taken from a Kaggle competition, there are a small number of submissions on the site of the challenge. These submissions are difficult to assess, as their evaluation at the time of writing is performed on only a subset of the test data, and using different criteria. The results in Table IV show that our model performs comparably in terms of prediction accuracy with the work presented for COVID-NET CT-1 and COVID-Net CT-2S/L. On reading the works by [10], [11], it is observed that the COVID-Net CT-1 model's evaluation score on the test dataset dropped from 99.1% to 94.5% when switching from COVIDx-CT to COVIDx-CT 2. The main difference between the datasets is that COVIDx-CT is based on images from Chinese provinces only, whilst the latter dataset is crowdsourced from a number of public international datasets.

We have also included the recently published work in [19] in the tabulated scores, as at the time of writing, it appears to be the current state-of-the-art approach evaluated on the COVIDx-CT 2A dataset. The approach in [19] is interesting as it re-introduced pretraining the CNN model on the ImageNet dataset prior to learning using the CT images. A final note is that our binary classification experiment of COVID-19 pneumonia against the normal controls discussed above is not featured in any of the tabulated works, where we obtained a best test accuracy of 97.09%. This result also confirms the flexibility of the model that can be easily adapted to tackle different problems.

V. Conclusions

There are a number of lessons learned from this study. One of the main observations from the conducted experiments was the slight improvement in model generalization for the binary classification. Considering the possibility that this was in part due to the contribution of the increased training sample sizes, it could be beneficial to the overall model scores to employ the data augmentation and preprocessing employed in [11] to further increase the dataset's robustness. The latter work included additional methods of data augmentation which were not considered in this work to avoid making modifications to the input data which would result in losing defining features, such as texture and shape.

Furthermore, the additional preprocessing used in [11] to omit visual indicators from outside of the patient's body could also improve prediction results. As for potential future work, the models presented in COVID-Net CT-1 and 2S/L are very portable. It could be a beneficial experiment to attempt to reduce the parameters of our approach and explore macro-architecture experiments as shown in this work. Moreover, the study in [19] also presented some interesting possibilities in terms of parameter initialization strategies, in addition to pre-training on ImageNet, both of which where attributed as contributors to their model's performance. There could also be potential in re-tuning the model's architecture to have convolutional layers near the final classification layer, which would make Grad-CAM a viable explainability tool.

In conclusion, this study presented a convolutional neural network capable of obtaining positive scores when predicting whether a chest CT scan belongs to patients with normal controls, common pneumonia, or COVID-19 pneumonia. The evaluation criteria used for this study also showed that the model obtained positive results in terms of false positives and false negatives. We believe that with further improvement, systems such as the work presented in this study can be used in a clinical context to assist with providing recommendations to medical experts during patient diagnosis.

References

[1] A. Tahamtan and A. Ardebili, "Real-time rt-pcr in covid-19 detection: issues affecting the results," *Expert review of molecular diagnostics*, vol. 20, no. 5, pp. 453–454, 2020.

[2] F. Pesapane, M. Codari, and F. Sardanelli, "Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine," *European radiology experimental*, vol. 2, no. 1, pp. 1–10, 2018.

[3] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of covid-19 in x-rays using ncovnet," *Chaos, Solitons & Fractals*, vol. 138, p. 109944, 2020.

[4] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with covid-19," *Nature medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[5] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans," *Nature Machine Intelligence*, vol. 3, no. 3, pp. 199–217, 2021.

[6] J. D. López-Cabrera, R. Orozco-Morales, J. A. Portal-Diaz, O. Lovelle-Enríquez, and M. Pérez-Díaz, "Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging," *Health and Technology*, vol. 11, no. 2, pp. 411–424, 2021.

[7] M. Perez Diaz, J. D. López Cabrera, O. Lovelle Enríquez, J. A. Portal Díaz, and R. Orozco Morales, "Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). the shortcut learning problem," 2021.

[8] A. Rehman, M. A. Iqbal, H. Xing, and I. Ahmed, "Covid-19 detection empowered with machine learning and deep learning techniques: A systematic review," *Applied Sciences*, vol. 11, no. 8, p. 3414, 2021.

[9] H. Gunraj, "Covidx ct," 2021, testing. [Online]. Available: https://www.kaggle.com/hgunraj/covidxct

[10] H. Gunraj, L. Wang, and A. Wong, "Covidnet-ct: a tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images," *Frontiers in Medicine*, vol. 7, p. 1025, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmed.2020.608525

[11] H. Gunraj, A. Sabri, D. Koff, and A. Wong, "Covid-net ct-2: enhanced deep neural networks for detection of covid-19 from chest ct images through bigger, more diverse learning," Jan 2021. [Online]. Available: https://arxiv.org/abs/2101.07433

[12] K. Team, "Keras documentation: 3d image classification from ct scans," Sep 2020. [Online]. Available: https://keras.io/examples/vision/3D_image_classification

[13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

[18] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[19] W. Zhao, W. Jiang, and X. Qiu, "Deep learning for covid-19 detection based on ct images," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.