Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Learning from imbalanced COVID-19 chest X-ray (CXR) medical imaging data

Jonathan H. Chan\textsuperscript{a,}\textsuperscript{*}, Chenqi Li\textsuperscript{b}

\textsuperscript{a} Innovative Cognitive Computing (IC2) Research Center, School of Information Technology, King Mongkut’s University of Technology Thonburi, Bangkok, Thailand
\textsuperscript{b} Division of Engineering Science, Faculty of Applied Science and Engineering, University of Toronto, Toronto, Canada

\textbf{ARTICLE INFO}

\textbf{Keywords:}
COVID-19
Chest X-ray
Medical imaging
Imbalanced data
Transfer learning
Deep neural networks

\textbf{ABSTRACT}

The trendy task of digital medical image analysis has been continually evolving. It has been an area of prominent and growing importance from both research and deployment perspectives. Nonetheless, it is necessary to realize that the use of algorithms, methodology, as well as the source of medical image data, must be strictly scrutinized. As the COVID-19 pandemic has been gripping much of the world recently, there has been much efforts gone into developing affordable testing for the masses, and it has been shown that the established and widely available chest X-rays (CXR) images may be used as a screening criteria for assistive diagnosis purpose. Thanks to the dedicated work by various individuals and organizations, publicly available CXR of COVID-19 subjects are available for analytic usage. We have also provided a publicly available CXR dataset on the Kaggle platform. As a case study, this paper presents a systematic approach to learn from a typically imbalanced set of CXR images, which consists of a limited number of publicly available COVID-19 images. Our results show that we are able to outperform the top finishers in a related Kaggle multi-class CXR challenge. The proposed methodology should be able to help guide medical personnel in obtaining a robust diagnosis model to discern COVID-19 from other conditions confidently.

1. Introduction

Fast, accurate and non-invasive method of COVID-19 (coronavirus disease 2019) diagnosis is a key to increase testing capacity, in order to minimize further community transmission. Currently, the internationally recognized and accepted method by the World Health Organization (WHO) is RT-PCR test with nasopharyngeal, nasal or throat swab \cite{1}. Such tests are invasive and can lead to side effects such as nose-bleed, headache, ear-ache, etc. Perhaps more importantly, the RT-PCR response times are typically at least two days, which leads to backlogs during pandemic peaks \cite{2}. This was observed in early cases from China \cite{3,4} and in various places in Europe \cite{5–7}. Although the use of chest CT (computed tomography) scans was common initially, these diagnoses have low specificity \cite{3}. The web annex on imaging for COVID-19 prefers the use of CXR over CT scans for COVID-19 diagnosis \cite{8}.

CXR images have been used to diagnose pneumonia with high accuracy, taking advantage of the recent advances in deep learning neural networks. Most recent advances have achieved 98.43\% accuracy \cite{9}. This inspires a similar application of COVID-19 classification using CXR images. However, the availability of high-quality COVID-19 CXR that has been properly diagnosed is still very limited. In fact, it has been argued that many publicly available “toy” datasets do not comply with clinical standards and a well-curated CXR dataset needs to involve professional radiologists and should be undertaking in multiple phases that span over one week’s time \cite{10}. Otherwise, the “garbage-in, garbage-out” concept means that the obtained diagnostic CXR models, however high accuracy, are unreliable.

This work takes an initial adhoc study and turned it into a systematic process to gather data, step-by-step analysis, including preprocessing, and eventual model-building with validation that would be suitable for the medical community.

2. Materials and methods

This study consists of relevant literature reviews to gather the current state-of-the-art prior works in imbalanced data treatment and deep learning analysis of medical images, data collection of CXR images from publicly available resources and quality assessment, development of robust deep neural network models of multi-class CXR images of COVID-19, and recommendation of a framework to deal with imbalanced data.

\textsuperscript{*} Corresponding author.
E-mail addresses: jonathan@sit.kmutt.ac.th (J.H. Chan), chenqi.li@mail.utoronto.ca (C. Li).

https://doi.org/10.1016/j.ymeth.2021.06.002
Received 2 February 2021; Received in revised form 7 May 2021; Accepted 1 June 2021
Available online 4 June 2021
1046-2023/© 2021 Elsevier Inc. All rights reserved.
that are commonly found in the real world. No ethical approval was
needed as the datasets were obtained from publicly available sources.
More details are provided in the following subsections.

2.1. Dataset preparation

2.1.1. Dataset collection

The following are the sources and related publications for the
data manually collected in this study. A mixed dataset of publicly available
COVID-19 CXR images from various sources 1, a pneumonia dataset
obtained from a study on children 2, partially verified thorax CXR im-
ages of adults from NIH (National Institute of Health, USA) 3, and
unverified no-finding images from NIH. The CXR images were curated
into three categories of NOFINDING, THORAXDISEASE, and COVID-19.
The NOFINDING case means the images are not associated with pneu-
monia. The dataset has been deposited as a Kaggle dataset [11]. The
COVID-19 images were semi-verified by the guidelines provided on a
medical-focus web blog.4

Training Dataset. The CXR dataset used composed of 363 COVID-19
images, 1,408 no finding images, 3,736 thorax disease (non-COVID-
19) images. They come from a multitude of sources as described above
and posted in a Kaggle classroom challenge.5

Leftout Dataset. Since training dataset employed downsampling using
the entire COVID-19 data, the remaining NOFINDING and THOR-
AXDISEASE images make up the leftout dataset.

Validation Dataset/Real World Unseen Data. The 1,130 unlabelled
CXR images used for evaluating model’s performance are unseen data
collected manually. These are not included in Version 1 of the Kaggle
CXR multi-class dataset, but they are available as an unlabelled dataset
at the Kaggle challenge.

2.1.2. Dataset sampling

Due to significant imbalance of the dataset, 10 times more thorax
disease images than COVID-19 images, the model to be trained can
easily overfit to the majority class. Downsampling, or known as under-
sampling, was applied to randomly choose 363 images from each of
the three classes. This base case is denoted as 1:1:1 for COVID-19:
THORAXDISEASE:NOFINDING. In order to investigate the effect of
imbalance data further, runs with ratios of 1:2:2 and 1:3:3 were also
conducted. More detailed information is given in the methods
subsection.

2.1.3. Dataset preprocessing and augmentation

Normalization of dataset by maximum pixel value of 255 was per-
formed. Image augmentation techniques such as rotation, horizontal
and vertical shift, ZCA whitening, zoom and shear were employed. They
will be further discussed in the methodology section as they potentially
have profound impact on the performance of the models trained.

2.2. Methods

2.2.1. Imbalanced dataset treatment

There has been various works in the area of class-imbalanced data
treatment, ranging from more traditional machine learning approaches
to recent extensions into the use of deep learning techniques, covering
data from management science to engineering [12]. Two basic ap-
proaches to address the data distribution imbalance are random over-
sampling of the minority classes and random undersampling of the
majority classes. A systematic study of the problems encountered for
treatments using convolutional neural networks found that over-
sampling is generally more effective compared with undersampling.
However, undersampling may be effective depending on the ratio and
extent of imbalance [13]. In addition, the effect of overfitting is reduced
with the use of convolutional neural network (CNN) models. This effect
of not needing to deal with imbalance in data vigorously was also found
in a study to predict the effect of genetic variants on gene splicing using
CNN [14]. In another recent study, the use of deep conditional generative
models has been shown to perform better than traditional over-
sampling in many cases, especially for severely imbalanced ones.
Although the effect of imbalanced data may be countered, the class
overlap problem may be more detrimental and difficult to handle [15].
As a mean to deal with highly imbalance big data, sampling and
thresholding strategies are shown to be effective when applying deep
learning techniques [16].

Based on the above premises, oversampling may seem to be a
reasonable approach. However, with the advent of big data, it would be
computational expensive to use oversampling excessively. Given the size
of medical images and the common practice to increase the resolution
for better performance, the oversampling approach is not very scalable.
A common approach is ensemble learning by undersampling numerous
times and use an ensemble classifier. While this has merits and can
improve the representation of the multi-classes, model deployment may
be more complex as it involves storing and executing numerous large
models. In order to address the above limitations, our approach is a
single undersampling model approach that uses the remaining data that
has not been used for training as a pseudo-test set to help in selecting a
suitable model for deployment. This turns out to be a simpler version of
the use of sampling and thresholding strategies mentioned earlier.

2.2.2. Deep learning techniques

Recent state-of-the-art advances in the field include CoroNet, a
transfer learning model based on the Xception architecture, with 54,528
trainable parameters. CoroNet achieved 95% 3-class classification ac-
curacy (COVID vs pneumonia vs normal) [17]. CovidGan, an Auxiliary
Classifier Generative Adversarial Network, helped produce synthetic
images to enhance performance CNN for COVID detection from 85% to
95% [18]. DarkCovidNet, which has a 17-convolutional layer network
with different filtering on each layer, achieved 87.02% for multi-class
classification (COVID vs no-findings vs pneumonia) [19]. Apostolo-
poulos et al. utilized transfer learning based on models such as VGG19,
MobileNet v2, Inception, Xception, Inception ResNet v2 to achieve
92–93% 3-class classification accuracy (COVID vs normal vs pneu-
monia) [20]. Hassan et al. applied XGB-L, XGB-Tree, CART, KNN
and Naïve Bayes to achieve accuracies ranging from 66.27% to 79.52% [21].
COVIDDiagnosis-Net, a Bayesian optimization-based SqueezeNet model,
pretrained on ImageNet dataset, achieved 98.26% accuracy on 3-class
classification (COVID vs normal vs pneumonia) [22]. DeTrAC tech-
nique composed of class decomposition, transfer learning and class
composition to achieve 93.1% 2-class classification accuracy [23].
COVIDPred combined 2 best-performing models from 29 different types
of models. They found that images rotated through 120 degrees or 140
degree angle displayed highest validation accuracy of 81.2% [24]. Chen
et al. compared different transfer learning base models, including

---

1 Cropped COVID-19 + selected images from the following (retrieved July 17, 2020): https://github.com/achung/Actualmed-COVID-chestxray-dataset
2 Pneumonia CXR images of children retrieved from Mendeley at: https://data.mendeley.com/datasets/rcbjbr9j/2
3 Selected images from: https://nihcc.app.box.com/v/ChestXray-NIHCC
4 More information can be found at: https://www.kaggle.com/c/dlai3- phases/
were imported and supplemented with one dropout layer and 2 dense
learning the peculiarities of dataset rather than the pathology [27].

Hasan et al. trained deep learning architecture based on ResNet and Xception to find potential issues of overfitting, bias and limitations of datasets commonly employed by researchers. They found that the high accuracy reported in recent algorithms are likely due to bias in experimental design and overfitting, namely cross-validation results without independent test set can overestimate network performance. A major issue was that data from different classes come from separate sources, network might focus on learning the peculiarities of dataset rather than the pathology [27].

From the prior arts mentioned above, the more successful architectures were based on Xception, Inception, Resnet or hybrid combinations of these. While there has been much work on trying to optimize the hyperparameters in order to improve the predictive performance of these models, it is a very time-consuming process, even with the advent of autoML (autonomous machine learning). However, given an unknown scenario, how well will these models perform? In addition, with big data fast becoming a norm, as well as most cases having imbalanced data, it is important to deal with imbalanced datasets, as addressed in the previous subsection.

In summary, much success has been found using oversampling when treating imbalanced cases, but the big data effect has forced others to look into undersampling more now. This work proposes a systematic approach to undersample multiclass imbalanced medical imaging data.

2.2.3. Base model selection

Simple CNN Verification. A simple convolutional neural network was able to achieve great performance for binary class classification of COVID-19 and non-COVID-19. However, the same model architecture was trained for the desired 3 class classification, but only 65–80% F1 score can be obtained compared to 98% F1 score for binary classification. As a result, it suggests the need for a more complex architecture, such as employing transfer learning using pretrained models as the starting point.

Transfer Learning Architectures. At the time of model selection, some of the most common state-of-the-art COVID-19 CXR classification algorithms use VGG19, DenseNet-201 or Xception as base module for transfer learning. In order to select the base module, these architectures were imported and supplemented with one dropout layer and 2 dense layer at the end. Training was performed separately on the aforementioned architecture. Amongst the three modules, Xception achieved the best F1 score of 86%. This choice is further supported by the state-of-the-art CoroNet, an Xception based architecture which achieved 95% 3-class classification.

2.2.4. Hyperparameter tuning

The hyperparameters considered include input image size, epochs, batch size, learning rate, image augmentation techniques (rotation, translation, ZCA whitening, zoom, shear). Given the large set of hyperparameters to tune, the number of possible combinations to be tested greatly exceeds the given computational power. Therefore, an evolutionary technique was employed based on survival of the fittest.

Survival of the Fittest. Proposed by Darwin’s evolutionary theory, the concept is based on the idea that life forms with the most desirable traits for the environment will succeed and continue to reproduce. Since neural networks itself is a mimicry of life itself, the choice of neural network architecture can also inherit a similar ideology. As a result, using CoroNet as the ancestor of all models, offsprings were reproduced with different values for a chosen hyperparameter. After evaluating the performance, the offspring that produced best result was chosen to be the parent for future generations. The process repeats until all hyperparameters were chosen. (See Fig. 1.)

2.2.5. Metrics

The performance metrics used were accuracy, precision, recall and F1. These may be calculated using the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) values obtained for each model. The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other. As this work concerns imbalanced multi-classes, the macro F1 average that gives the same weight to each class is more suitable. The F1 results from classification given in tabular form would refer to macro F1 values.

Accuracy. Accuracy is the ratio of correct predictions (TP + TN) to all predictions (TP + TN + FP + FN)

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

Precision. Precision is the ratio of true positives (TP) to all predicted positives (TP + FP)

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

Recall. Recall is the ratio of true positives (TP) to all actual positives (TP + FN)

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

F1. F1 score combines the statistical measures of precision (p) and recall (r)

\[
F1 = 2 \frac{p \cdot r}{p + r}
\]

2.2.6. Proposed systematic undersampling approach

One challenge faced by the evolution-inspired hyperparameter tuning workflow is choosing the best performing model among the same generation. It is known that performance metrics of model based on the training set alone is not a good indication of its performance in the real world due to potential overfitting or underfitting.

The approach to solving this problem is to take advantage of the leftout data due to downsampling of the original dataset.

In an attempt to investigate the relationship between model’s performance evaluated against training set, leftout set, and validation set, a model architecture was selected and many training instances were created. Each training instance saved three model callbacks: model with highest accuracy (best accuracy model), model with lowest loss (best loss model) and model after all training epochs have elapsed (completed model). Each of the saved models is then evaluated using the training, leftout and validation sets. The results are summarized in Fig. 2, 3 and 4 in the next section, with each color representing one training instance.

3. Results and discussion

3.1. Best performing model

Using the survival of the fittest hyperparameter tuning technique, the best performing model (seed value 41) in the initial trial-and-error experiments was able to achieve an overall 92.58% macro F1 score on the 1,130 images of previously unseen real world dataset. When submitting this result to the Kaggle server, the private leaderboard macro F1 score was 93.05%, which outperformed the 92.05% score of the top finisher in the Kaggle competition. The random seed value for this run was 41.

Then to assess the reproducibility of the model runs, five repeated
runs at the same conditions but using a random seed were done sequentially. These results showed significant stochasticity in the algorithm used. To minimize this effect, a seed value was set with the random undersampling process for all subsequent runs. The seed value that provided a good representative downsampling of the given training dataset was 41.

To further validate the model and mitigate the randomness of stochastic gradient descent, twenty repeated runs were performed using the seed value of 41. The averaged results are shown in Table 1. The macro weighted average F1 score was 89.19%. The best macro F1 score of these twenty runs was 93.33%, and the corresponding private kaggle leaderboard macro F1 score was 93.18%.
3.1. Model architecture

The best performing model architecture and hyperparameters are summarized in Tables 2 and 3. Even though 160 epochs generated the best results in the initial studies, the repeated runs were performed with 100 epochs as the performance began to level off at around that point.

3.2. Correlation between model performance on train, leftout and real datasets

Using the aforementioned model architecture, two different sets of experiments were performed to investigate the correlation between model’s performance on training, leftout, and real world unseen dataset. In the exploratory experiment, 16 training instances with random seeds were created. The results showed that recall and F1 have moderate correlation between the leftout and the real world unseen dataset. However, for the precision measure, the opposite was found that the training data correlated with the unseen data instead. This illustrates the danger that the split between the train and leftout images of any category, in this case NOFINDING, may not be appropriately distributed for many of the 16 random seed runs. Therefore, it is necessary to search for a representative distribution of the training and leftout datasets. This was done by repeated random sampling with different seeds. For our dataset, a seed value of 41 was found to be appropriate. Results from twenty repeated runs with seed 41 are summarized in more detail in the following subsections.

3.2.1. Recall

The recall measure is often the most important in medical settings. In the case of COVID-19, the model’s ability to correctly identify all COVID-19 patients, with minimal false negative is strongly desirable, as the model can serve as a preliminary test to quickly identify potential patients and prevent further spread. Therefore, an accurate estimation of model’s recall on real world data will be a helpful tool to select the most appropriate model. In particular, model’s recall evaluated using leftout set was often a precise indicator of model’s recall on real world data, as can be seen qualitatively in Fig. 2. As can be observed from the results, the train measure often results in much overfitting. That is, the recall scores are much lower in the real data. However, the leftout recall scores are more similar in magnitude as the real ones and show more similar trends as well. As a rule-of-thumb based on these results, the train recall value should be greater than 0.96 and the corresponding leftout recall value should be over 0.90, based on the best accuracy/loss model, in order to have a robust model that can generalize well.

3.2.2. F1 score

Similar to the recall score, the leftout set was also found to be a decent indicator of the expected performance of the model on real world data, as can be seen in Fig. 3. These are macro F1 values that give the same importance to the minority COVID-19 class as the much larger NOFINDING and THORAXDISEASE classes. The same rule-of-thumb value for recall may be applied for the F1 score, using the best accuracy/loss model.

3.2.3. Precision

Precision has a weaker correlation between the leftout set and real world data from this experiment, as can be seen in Fig. 4. This indicates that the real world unseen dataset may have a different distribution than the training set and may contain unseen patterns. Nonetheless, this is an important measure for the minority COVID-19 class. Thus, an addition rule of a minimum precision value of COVID-19 class should be implemented. This value would depend on the comfort level of the medical facility. Typically, a train value of at least 0.95 may be required.
3.2.4. Correlation analysis

To further support the visualization, scores were plotted for train vs real and leftout vs real, and the corresponding correlation coefficients were calculated. This provides a quantitative assessment of the results. As can be seen in Fig. 5, for recall, F1 and precision, leftout vs real had a much higher correlation score compared to train vs real. This is an indication that model’s performance on real world data can be inferred from its performance on leftout dataset. In particular, for recall and F1, the correlation coefficient was 0.69, 0.64 and 0.50 respectively for leftout vs real, comparing that to 0.31, 0.32 and 0.22 respectively for train vs real.

The quantitative results suggest that there is a moderate correlation between leftout and real dataset for recall and F1, and a marginal correlation for precision. The rule-of-thumb guideline from the qualitative analyses provides a more practical mean for implementation, with the correlation results supporting the use of leftout dataset for both the recall and F1 measures.

We would like to point out that the aforementioned results are observations based on COVID-19 CXR classification. While similar results might be expected in other CXR classification tasks, perhaps even in other medical and non-medical contexts, no results are available to support such claims. Next steps include performing similar correlation analysis on other CXR datasets.

3.3. Imbalanced dataset treatment

The aforementioned results are obtained using a balanced dataset with 1:1:1 ratio for each of the three classes. Following a similar approach to prior works that were cited in an earlier section, we adjusted the undersampling ratio of the other two categories to be 2 and 3 times that of the minority COVID-19 category. That is we have ratios of 1:1:1, 1:2:2, and 1:3:3 number of points for COVID19:THORAXDISEASE:
Fig. 5. Correlation Analysis for Recall, F1 and Precision.
NOFINDING. We run each ratio 20 times to obtain more representative sources to facilitate research efforts in improving the robustness of CXR would not overfit excessively. If resources permit, should obtain a model on recall and F1 metrics can be used to identify robust models that dataset experienced degradation of performance on hold-out data, whereas models trained using imbalanced dataset did not show such degradation and sometimes even improved the result. However, the training time increased proportionally going from 1:1:1 to 1:2:2 to 1:3:3.

In comparison to the Kaggle top result of 92.05% F1 score in the private leaderboard were 93.18%, 92.76%, and 93.57% from global competitors. Finally, we proposed a framework to deal with novel, imbalanced data, in particular, using the leftout dataset as a pseudo-test set to get a better grasp of the model performance on real world data.

4. Conclusion

In this work, a three-class dataset was composed from a multitude of sources to facilitate research efforts in improving the robustness of CXR classifiers for COVID-19 classification. Furthermore, through the use of a novel, "survival of the fittest", hyperparameter tuning approach, we were able to propose a deep neural network model based on balanced undersampling that is capable of achieving an overall F1 score of 93.33% on previously unseen real world data. This outperforms the top finisher in the multi-class CXR COVID-19 Kaggle classroom challenge from global competitors. Finally, we proposed a framework to deal with imbalanced data, in particular, using the leftout dataset as a pseudo-test set to get a better grasp of the model’s performance on real world data.

4.1. Future work

An improvement to enhance the reproducibility of models is to set a seed for every random operation. The random initialization of the network and downsampling of the dataset can greatly vary model’s performance.

As discussed in the work of Xiao et al., radiologists often assess opacification severity zones in segmented CXR, in order to estimate if and when intubation or ventilation unit would be required. Therefore, a future extension to the model would leverage domain experts such as radiologists to highlight the severity zones and provide score prediction to help doctors assess patient’s severity and how rapidly the condition may deteriorate.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The second author would like to acknowledge stipend support for his summer research internship from ESROP Program of the Division of Engineering Science, Faculty of Applied Science and Engineering, University of Toronto, and School of Information Technology, King Mongkut’s University of Technology Thonburi.

References

[1] E.A. Akl, I. Blazic, S. Yaacoub, G. Frija, R. Chou, J.A. Appiah, M. Fatemi, N. Flor, E. Hitti, H. Jafri, J. Ifti, Z.-Y. Jin, H.U. Kuznezov, M. Kawyouna, E.A. Kazerooni, J.P. Ko, R. Mahfouz, V. Muglia, R. Nyabanda, M. Sanchez, P.B. Shete, M. Ulla, C. Zheng, E. van Deventer, M. d. R. Perez, Use of chest imaging in the diagnosis and management of covid-19: A WHO rapid advice guide, Radiology 298 (2) (2021) E63-E99, PMID: 32929811. doi:10.1148/radiol.20202003173. URL; http://www.roi.ai/article/10.1148/radiol.2020190011.

[2] A. Cozzi, S. Schiaffino, F. Arpaia, G. Della Pesa, S. Triteila, P. Bertolotti, L. Menicagui, C.G. Monaco, L.A. Carbonaro, R. Spairani, B. Babaei Paskef, F. Sardanelli, Chest x-ray in the COVID-19 pandemic: Radiologists' real-world reader performance, European journal of radiology 132 (2020) 109272–109272, edition: 2020/09/10 Publisher: Elsevier B.V. doi:10.1016/j.ejrad.2020.109272. https://pubmed.ncbi.nlm.nih.gov/32971526.

[3] H. Kim, H. Hong, S.H. Yoon, Diagnostic Performance of CT and Reverse Transcriptase Polymerase Chain Reaction for Coronavirus Disease 2019: A Meta-Analysis, Radiology 296 (3) (2020) E145–E155, edition: 2020/04/17 Publisher: Radiological Society of North America. doi:10.1148/radiol.2020201343. https://pubmed.ncbi.nlm.nih.gov/32301646.

[4] V. Zhao, C. Xiang, S. Wang, C. Peng, Q. Zou, J. Hu, Radiology department strategies to protect radiologists against COVID19: Experience from Wuhan, European journal of radiology 127 (2020) 108896–108896, edition: 2020/04/20 Publisher: Elsevier B.V. doi:10.1016/j.ejrad.2020.108896. https://pubmed.ncbi.nlm.nih.gov/32344294.

[5] S. Koorkai, M. Hosseiny, L. Myers, A. GholamrezaNezhad, Coronavirus (COVID-19) Outbreak: What the Department of Radiology Should Know., Journal of the American College of Radiology: JACR 17 (4) (2020) 447–451. doi:10.1016/j.jacr.2020.02.008.

[6] N. Flor, R. Dore, F. Sardanelli, On the Role of Chest Radiography and CT in the Coronavirus Disease (COVID-19) Pandemic., AJR. American journal of roentgenology 215 (4) (2020) W4,4, edition: United States. doi:10.2214/ AJR.20.23411.

[7] M. Zanardo, S. Schiaffino, F. Sardanelli, Bringing radiology to patient’s home using mobile equipment: A weapon to fight covid-19 pandemic, Clinical Imaging 68 (2020) 99-101, doi: 10.1016/j.clinimag.2020.06.031. http://www.sciencedirect.com/science/article/pii/S0957417416307175.

[8] R. Chou, M. Pappas, D. Buckley, M. McDonagh, A. Totten, N. Flor, F. Sardanelli, T. Dana, E. Hilt, N. Wasson, H. Nelson, Use of chest imaging in covid-19: a rapid advice guide.

[9] M.F. Hashmi, S. Kitayar, A.G. Keskar, N.D. Boldke, Z.W. Geem, Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning, Diagnostics (Basel, Switzerland) 10 (6) (2020) 417, publisher: MDPI. doi:10.3390/diagnostics10060417. https://pubmed.ncbi.nlm.nih.gov/32575475.

[10] H.R. Tizhoosh, J. Fratesi, COVID-19, AI enthusiasts, and toy datasets: radiology diagnostics10060417. https://pubmed.ncbi.nlm.nih.gov/32575475.

[11] J.H. Chan, Dlai3 hackathon phase3 covid-19 cxe challenge. kaggle. doi:10.34740/DLAI3-2020.23411.

[12] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanaye, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Systems with Applications 73 (2017) 220–239, doi: 10.1016/j.eswa.2016.12.035. http://www.sciencedirect.com/science/article/pii/S0957417416307175.

[13] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, CoRR abs/2003.00356.

[14] J.H. Chan, Dlai3 hackathon phase3 covid-19 cxe challenge. kaggle. doi:10.34740/DLAI3-2020.23411.

[15] P. Vuttipitayamonkong, E. Elyan, A. Petroskiv, On the class overlap problem in imbalanced data classification, Knowledge-Based Systems 212 (2021), 106631, doi: 10.1016/j.knosys.2020.106631. URL; http://www.sciencedirect.com/science/article/pii/S0957417416307175.

[16] J. Johnson, T. Khoshgoftaar, Thresholding strategies for deep learning with highly imbalanced big data., Deep Learning Applications, Volume 2. Advances in Intelligent Systems and Computing, vol 1232.In: Wani M.A., Khoshgoftaar T.M., editors. Springer, Singapore. doi: 10.1007/978-981-15-65759-9_9.

[17] A.J. Khan, J.I. Shah, M.M. Bhat, GuroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, Computer methods and programs in biomedicine 196 (2020) 105811–105851, edition: 2020/06/05 Publisher: Elsevier B.V. doi:10.1016/j.cmpb.2020.105851. https://pubmed.ncbi.nlm.nih.gov/32534344.

[18] A. Wahed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P.R. Pinheiro, Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection, IEEE Access 8 (2020) 19166–19192, doi: 10.1109/ACCESS.2020.2949762.

[19] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U. Rajendra Acharya, Automated detection of covid-19 cases using deep neural networks with x-ray images, Computers In Biomed. Res. 121 (2020), 103792, doi:10.1016/j.combrep.2020.103792. URL: http://www.sciencedirect.com/science/article/pii/S0010482520301621.

[20] I.D. Apostolopoulos, T.A. Mpastasa, Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks, Physical and
[21] L. Hussain, T. Nguyen, H. Li, A.A. Abbasi, K.J. Lone, Z. Zhao, M. Zaib, A. Chen, T. Q. Duong, Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection, BioMedical Eng. OnLine 19 (1) (2020) 88, https://doi.org/10.1186/s12938-020-00831-x.

[22] F. Ucar, D. Korkmaz, Covidiagnosis-net: Deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (covid-19) from x-ray images, Medical Hypotheses 140 (2020), 109761, https://doi.org/10.1016/j.mehy.2020.109761. URL: http://www.sciencedirect.com/science/article/pii/S0306987720307702.

[23] A. Abbas, M.M. Abdelsamea, M.M. Gaber, Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network, Appl. Intelligence 51 (2) (2021) 854–864, https://doi.org/10.1007/s10489-020-01829-7.

[24] A. Sharma, S. Rani, D. Gupta, Artificial Intelligence-Based Classification of Chest X-Ray Images into COVID-19 and Other Infectious Diseases, Int. J. Biomed. Imaging 2020 (2020) 8889023, publisher: Hindawi. doi:10.1155/2020/8889023. doi: 10.1155/2020/8889023.

[25] A. Chen, J. Jaegerman, D. Matic, H. Inayatali, N. Charoenkitkarn, J. Chan, Detecting covid-19 in chest x-rays using transfer learning with vgg16, in: CSBio ’20 Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics, Association for Computing Machinery, New York, NY, USA, 2020, pp. 93–96, https://doi.org/10.1145/3429210.3429213.

[26] C. Li, M. Wang, G. Wu, K. Rana, N. Charoenkitkarn, J. Chan, Covid19 chest x-ray classification with simple convolutional neural network, in: CSBio ’20 Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics, Association for Computing Machinery, New York, NY, USA, 2020, pp. 97–100, https://doi.org/10.1145/3429210.3429216.

[27] M.K. Hasan, M.A. Alam, M.T.E. Elahi, S. Roy, S.R. Wahid, R. Martin, B. Khanal, Challenges of deep learning methods for covid-19 detection using public datasets, medRxiv doi:10.1101/2020.11.07.20227504. https://www.medrxiv.org/content/early/2020/11/10/20227504.

[28] N. Xiao, J.G. Cooper, J.M. Godbe, M.A. Bechel, M.B. Scott, E. Nguyen, D. M. McCarthy, S. Abboud, B.D. Allen, N.D. Parekh, Chest radiograph at admission predicts early intubation among inpatient covid-19 patients, European Radiol. (2020) 1–8.