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.
Performance change with the number of training data: A case study on the binary classification of COVID-19 chest X-ray by using convolutional neural networks

Kuniki Imagawa, Kohei Shiomoto
Tokyo City University, Faculty of Information Technology, 1-28-1 Tamaasutsumi, Setagaya-ku, Tokyo, 158-8557, Japan

ABSTRACT

One of the features of artificial intelligence/machine learning-based medical devices resides in their ability to learn from real-world data. However, obtaining a large number of training data in the early phase is difficult, and the device performance may change after their first introduction into the market. To introduce the safety and effectiveness of these devices into the market in a timely manner, an appropriate post-market performance change plan must be established at the timing of the premarket approval. In this work, we evaluate the performance change with the variation of the number of training data. Two publicly available datasets are used: one consisting of 4000 images for COVID-19 and another comprising 4000 images for Normal. The dataset was split into 7000 images for training and validation, also 1000 images for test. Furthermore, the training and validation data were selected as different 16 datasets. Two different convolutional neural networks, namely AlexNet and ResNet34, with and without a fine-tuning method were used to classify two image types. The area under the curve, sensitivity, and specificity were evaluated for each dataset. Our result shows that all performances were rapidly improved as the number of training data was increased and reached an equilibrium state. AlexNet outperformed ResNet34 when the number of images was small. The difference tended to decrease as the number of training data increased, and the fine-tuning method improved all performances. In conclusion, the appropriate model and method should be selected considering the intended performance and available number of data.

1. Introduction

The extensive interest in artificial intelligence (AI) and machine learning (ML) application is growing in the medical field. This is primarily driven by the impressive progress made by deep learning (DL) as a subset of ML because of the increased computational power and an explosion in the availability of large datasets. The number of research papers on the application of DL to medical fields has increased. For the medical image analysis, the number has dramatically increased since 2015. The number of academic papers in major conferences and journals exceeded 300 by the end of 2016 [1]. Furthermore, more than 60 AI/ML-based medical devices have already been approved by the U.S. Food and Drug Administration (FDA) in the United States [2]. There are several types of AI/ML-based medical devices. Muehlematter et al. [3] reported that most AI/ML-based medical devices approved by the FDA are used in radiology, but they span across various medical specialties, such as cardiovascular and neurology. For example, applications in radiology can be categorized into classification, detection, segmentation, etc., and the required performance varies greatly depending on the modality and target diseases.

AI/ML-based medical devices are roughly divided into two types: 1) locked type, which fixes performance prior to marketing and unable to change performance with use; and 2) continuous type, which can change performance by continuously training data after market introduction. To date, several FDA-approved AI/ML-based medical devices are typically locked type, but the FDA announced its marketing authorization for the continuous type on February 2020 [4]. The number of these medical devices is expected to increase in the market in the future. In April 2019, the FDA published a discussion paper for a proposed regulatory framework to account for the iterative nature of AI/ML-based medical devices [5]. The paper describes the total product lifecycle approach. As part of this framework, the necessity of submitting a predetermined change control plan, in which manufacturers are anticipated to perform modifications to performance, inputs, or intended use prior to marketing, is...
also described. Other jurisdictions are also preparing papers on regulatory guidance, with a concept similar to that in this discussion paper. Therefore, an appropriate post-market performance change plan must be established at the timing of the premarket approval.

Since the outbreak of the Coronavirus Disease 2019 (COVID-19), many studies have used convolutional neural networks (CNNs) to detect COVID-19 on chest X-ray (CXR) images. For the COVID-19 diagnosis, testing through viral RNA identification in reverse transcriptase polymerase chain reaction (RT-PCR) is currently recommended. However, chest imaging techniques, such as computed tomography (CT) and chest radiography, are considered as part of the diagnostic workup of patients with suspected or probable COVID-19 disease in case RT-PCR is not available or the results are delayed or initially negative in the presence of symptoms suggestive of COVID-19 [6,7]. Applying ML methods to COVID-19 radiological imaging may improve the diagnosis accuracy compared with the gold-standard RT-PCR while providing a valuable insight into the prognostication of patient outcomes. In particular, chest radiography is widely used, takes less imaging time, and has an accessible diagnostic modality that may be easily brought to the patient’s bed. Arising from the success of the ImageNet Large-scale Visual Recognition Challenge (ILSVRC) in 2012 [8], the CNN is a class of artificial neural networks that has become dominant in object recognition, including radiology [9]. Most of the recent papers on COVID-19 diagnosis were conducted based on existing off-the-shelf models and classified images into two [10–13] or three classes [13–15]: COVID-19 and Normal or COVID-19, non-COVID-19 pneumonia, and Normal. These performances such as accuracy achieves exceed 90% by using data augmentation, transfer learning or combining of publicly available datasets. For example, Nayak et al. [16] evaluated the performance of eight pre-trained fundamental CNN models, which achieved excellent results in the ILSVRC, to classify COVID-19 and Normal. They also conducted a comparative analysis by considering hyperparameters, such as batch size, learning rate, and optimizer. ResNet34 [17] exhibited the best performance, followed by AlexNet [18]. Rahaman et al. [19] evaluated the performance of 15 pretrained fundamental CNN models to classify COVID-19, non-COVID-19 pneumonia, and Normal. VGG 19 [20] obtained the highest classification accuracy. Tuan [21] performed two- and three-class tasks to classify COVID-19 using three fine-tuned CNN models (i.e., AlexNet, GoogleNet [22], and SqueezeNet [23]) without data augmentation and achieved a high classification performance in terms of accuracy, sensitivity, specificity, precision, F1 score, and area under the curve (AUC). The results suggested that the fine-tuning of network learning parameters is important because it can help avoid the development of more complex models when existing ones can achieve the same or much better performance. On the contrary, many current studies were conducted based on a small number of training data or a combination of training data without demographic statistics (e.g., age and sex distributions) due to the limited publicly available datasets. These estimated performances may probably be optimistic and misleading because of the high-risk bias based on the non-representative selection of control patients and model overfitting [24,25]. In 2020, Alberto et al. [26] publicly released a large and fully annotated BrixIA dataset of 4703 CXR images related to COVID-19 with additional information on severity, participants’ age and sex, and modality manufacturer. A large dataset from the Valencian Region Medical Image Bank containing 3141 CXR and 2239 CT images of patients with COVID-19 along with their radiological findings and locations, pathologies, radiological reports, Digital Imaging and Communications in Medicine (DICOM) metadata, diagnostic antibody tests, etc., was also publicly released [27]. Such large databases are expected to be actively developed and made available to the public in the future. It is hoped that these large datasets with patient demographics will enable the introduction of appropriate AI/ML-based medical devices to the market to support medical decision making through proper regulation.

With the abovementioned background, this study evaluates the performance change as the number of training data increases. Two different CNNs, namely AlexNet and ResNet34, with and without a fine-tuning method are used to classify COVID-19 and Normal. A large BrixIA dataset of CXR images related to COVID-19 are used as the training, validation, and test data. The AUC, sensitivity, and specificity are utilized as the performance evaluation items because they are mainly evaluated through FDA premarket approvals [28]. The major outcomes of this study are the following: 1) All performances were rapidly improved as the number of training data were increased and reached an equilibrium state. 2) AlexNet outperformed ResNet34 when training data were small, and the difference between the performance of AlexNet and ResNet34 decreased as the training data increased. 3) The fine-tuned CNNs performed better than CNNs trained from scratch in all training datasets; this effect is particularly noticeable in small training data. 4) The change in the performance of the binary classification for COVID-19 and Normal datasets can be generalized as COVID-19 and non-COVID-19 pneumonia datasets.

2. Material

2.1. Datasets

Two independent datasets were used for the evaluation. The first dataset is the BrixIA dataset, which comprises 4703 CXR images of COVID-19 objects taken for both triage and patient monitoring in sub-intensive and intensive care units for 1 month between March 4 and April 4, 2020 at ASST Spedali Civili di Brescia. The images were...
retrieved from the facility’s Picture Archiving and Communication Systems (PACS) and disclosed as DICOM formats with additional information (i.e., severity, patient’s age and gender, and modality manufacturer). The second data set is a Chest X-ray14 dataset comprising 112,120 CXR images with 14 diseases and one for the Normal label from 30,805 unique patients. The CXR images were extracted from the PACS database through natural language processing (NLP) at the National Institutes of Health Clinical Center between 1992 and 2015 and disclosed as portable network graphics formats with additional information (i.e., patient ID, age, and gender and view position).

The DICOM images in the BrixI dataset, whose window width (WW) and window center (WC) could not be derived from DICOM Tags, were excluded. The remaining DICOM images were converted from 16-bit into 8-bit through windowing, which changed the picture’s appearance to highlight particular structures using the WW and WC derived from the DICOM Tags. Fig. 1 depicts an example of the CXR images from both COVID-19 and Normal classes. We randomly selected 4000 COVID-19 images and 4000 CXR images labeled as Normal from the Chest X-ray14 dataset. The combined dataset, which contained 8000 CXR images, was split into two to separate 7000 images for training and validation and 1000 images for testing. The training and validation data were selected as 16 different datasets ($N = 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, and 7000$). The ratio of the COVID-19 and Normal classes had the same proportion in all the training, validation, and test data.

Table 1 and Fig. 2 depict the detailed patient demographics and age distribution for each dataset.

| Database Origin | Purpose               | Label      | Images | Patients | Female | Male | Age   | AP | PA |
|-----------------|-----------------------|------------|--------|----------|--------|------|-------|----|----|
| BrixI           | Training and Validation Test | COVID-19   | 3500   | 1804     | 1060   | 2440 | 59 ± 14 | –  | –  |
|                 |                       |            | 500    | 429      | 130    | 370  | 58 ± 14 | –  | –  |
| Chest X-ray 14  | Training and Validation Test | Normal     | 3500   | 3026     | 1521   | 1979 | 45 ± 17 | 1229 | 2271 |
|                 |                       |            | 500    | 491      | 223    | 277  | 45 ± 17 | 175 | 325 |

Table 1
Data and patient characteristics. AP: anteroposterior and PA: posteroanterior representing the view position. We cannot find AP and PA in the DICOM Tags, but the ratio is reported as AP (87%) and PA (13%) in the BrixI dataset.

3. Methodology

Figs. 3 and 4 show the proposed method for classifying COVID-19 and Normal, which mainly consisted of preprocessing and classification with a fine-tuned CNN. The detailed preprocessing, classification
and evaluation are described in the subsequent sections. The 16 training and validation datasets fed to each CNN model with and without a fine-tuning method were evaluated on the common test dataset.

3.1. Preprocessing

Before feeding the CXR images to the system as input, all CXR images were resized to $256 \times 256$ px and cropped in the center as $224 \times 224$ px. The grayscale images were converted into a colored format to three channels (RGB: red, green, and blue). The pixel values of the input images were normalized in between ranges 0 and 1 based on the mean and the standard deviation to maintain the numerical stability in the CNN architectures and use a pretrained CNN with ImageNet, which contains 1.4 million images with 1000 classes.

3.2. Classification

The AlexNet and ResNet34 models were used for the classification. AlexNet was treated as the first breakthrough in the CNN model architecture, serving as the winner of the 2012 ILSVRC. It adopted an eight-layer network structure consisting of five convolutional layers and three fully connected layers. After each convolution in the five convolutional layers, maximum pooling was performed to reduce the amount of data. Data augmentation and dropout were used to reduce overfitting. Rectified linear units were utilized as an activation function instead of a sigmoid or hyperbolic tangent function. AlexNet had eight layers, and the number of parameters was approximately 60 million. Meanwhile, ResNet was the winner of the 2015 ILSVRC, which enables the training of up to hundreds or thousands of layers and inspired many other models. When the network is too deep, the gradients from where the loss
function is calculated easily shrink to zero after several chain rule applications. This result on the weights never updates its values; therefore, no learning is performed. ResNet overcomes this degradation problem by introducing residual connections mapping to fit input from a previous layer to the next layer and achieves compelling performance. ResNet34 had 34 layers and 21.8 million parameters.

In addition to the AlexNet and ResNet34 CNN models without pretraining, we utilized a pretrained network, called the fine-tuning method. This method is often applied to radiology studies to replace the fully connected layers of the pretrained model with a new set of fully connected layers to retrain on a given dataset and finetune all the kernels in the pretrained convolutional base by means of backpropagation. All convolutional base layers can be finetuned. Alternatively, some earlier layers can be fixed while fine-tuning the remaining deeper layers. Our method requires the unfreezing of the entire model originally trained on a large-scale labeled dataset called ImageNet and re-training it on CXR images. Data augmentation was not conducted herein. This work is motivated by the observation that early-layer features appear more generic (e.g., edges applicable to various datasets and tasks), while later features progressively become more specific to a particular dataset or task [29,30].

3.3. Evaluation

The AUCs derived from the receiver operating characteristic (ROC) curve, sensitivity, and specificity were used for the evaluation. The ROC curve is graphical display of the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis for varying the cut-off points. Both axes are from 0 to 1. The TPR is equal to the sensitivity, while the FPR is equal to “(1 - specificity)”. The AUC is the definite integral of an ROC curve and an effective and combined measure of sensitivity and specificity that assesses the inherent validity of a diagnostic test. An AUC closer to 1 indicates a better test performance. Sensitivity and specificity were determined by using the cut-off point defined as the Youden Index [31]. The Youden Index is the point on the ROC curve that is farthest from the line of equality (diagonal line). The optimal cut-off value is that at which the determined “sensitivity + specificity - 1” is maximized. To construct a 95% confidence interval, the standard error was calculated using the Hanley and McNeil method [32] for the AUC and the Wald method [33] for sensitivity and specificity. A nonlinear function, \( y = a - bx^c \), was also fitted to each data. All results will be described herein to show the relation between each performance and the number of training images.

4. Experiment and results

The 16 training and validation datasets were divided into 32 batches. The training of the AlexNet and ResNet34 CNN models with and without a fine-tuning method was conducted by using an Adam optimizer (\( \beta_1 = 0.9, \beta_2 = 0.999 \)). The network was trained 50 epochs. The learning rate was set to \( 1 \times 10^{-5} \) for AlexNet and \( 1 \times 10^{-6} \) for ResNet34. The hyperparameters were determined based on the comprehensive study of Nayak et al. [16]. To assess the performance of each CNN model, the ROC curve and the AUC on the common test dataset were initially
determined through training on the 16 datasets. Fig. 5 shows the relations between the AUC and the number of training images for all CNN models. The sensitivity and the specificity were determined based on the cut-off values determined by using the Youden Index for each ROC curve. Fig. 6 presents an example of an ROC curve and the cut-off values obtained using the AlexNet model for the following number of CXR images: 500, 1000, 2000, and 4000 images. The sensitivity and the specificity for each ROC curve were determined. Figs. 7 and 8 depict the relationship between the sensitivity and the number of training images and between the specificity and the number of training images. Figs. 5, 7, and 8 show the 95% confidence interval and the fitted nonlinear function, in which parameters $a$, $b$, and $c$ were calculated using the open-source Python library used for scientific and technical computing (SciPy; version 1.4.1) with the Levenberg–Marquardt algorm. Table 2 presents parameters $a$, $b$, and $c$.

5. Discussion

One of the features of AI/ML-based medical devices resides in their ability to learn from real-world data. However, obtaining a large number of training data in the early phase is difficult, and the device performance may change after their first market introduction. To introduce the safety and effectiveness of these devices into the market in a timely manner, an appropriate post-market performance change plan should be established at the timing of the premarket approval, and the real-world performance change must be monitored. In this work, we studied how performance changes when the number of training data is changed. Fig. 9 shows the relations between the AUC and the training sample size and between the increment of the AUC and that of the training sample size by using the nonlinear function obtained from Table 2. All AUCs were rapidly improved as the training data increased and reached an equilibrium state. The 95% confidence interval also decreased as the training data increased. Fang et al. [34] and Samala et al. [35] obtained similar trends for deep learning-based organ auto-segmentation for
Fig. 8. Relation between specificity and the number of trained CXR images.

Table 2
Determined parameters of the nonlinear function: $y = a - bx^{-c}$.

| Model                     | AUC    | Sensitivity | Specificity |
|---------------------------|--------|-------------|-------------|
|                           | a      | b           | c           |
|                           | A      | A           | A           |
|                           | A      | A           | A           |
| AlexNet                   | 1.0072 | 1.1554      | 0.5274      |
| Fine-Tuned AlexNet        | 1.0033 | 0.0502      | 0.2715      |
| ResNet34                  | 0.9994 | 70.156      | 1.0889      |
| Fine-Tuned ResNet34       | 0.9991 | 78.6257     | 1.3567      |

Fig. 9. AUC versus the training sample size and the increment of the AUC versus the increment of the training sample size.
head-and-neck patients by using CT images and binary classification of malignant and benign masses in digital breast tomosynthesis. Whether the performance of the preapproval stage is in the process of a rapid change or in a steady state must be determined. If the performance is in the process of a rapid change, the manufacturer and regulatory authority should carefully monitor the real-world performance change after the first market introduction.

Alternatively, the AUCs for fine-tuned CNNs were better than those for CNNs trained from scratch in all training datasets. This effect is particularly noticeable when the training data are small. Tajbakhsh et al. [36] considered four distinct medical imaging applications in three specialties (radiology, cardiology, and gastroenterology) involving classification, detection, and segmentation under three imaging modalities and compared the performance of deep CNNs trained from scratch and fine-tuned pre-trained CNNs using ImageNet. They concluded that deeply fine-tuned CNNs are useful for medical image analysis, performing equally as CNNs trained from scratch and even outperforming them when limited training data are available. Furthermore, they indicated that the performance gap between deeply fine-tuned CNNs and those trained from scratch widened when the size of training sets was reduced. Our result agrees with their findings, which validates the generality of the mentioned case.

Most of the previous studies used publicly available datasets labeled as COVID-19, which are limited to tens to a few hundreds by using data augmentation, transfer learning or combining of datasets. Ran et al. [37] collected over 10 000 CXR images labeled COVID-19 and non-COVID-19 pneumonia from five hospitals and more than 30 clinics by using NLP. In their study, a binary classification was performed by using a modified DenseNet model. The AUC versus the training sample size and the increment of the AUC versus that of the training sample size was reduced. Our result agrees with their findings, which confirmed. Their results demonstrated that more than 3000 training samples are needed to achieve an AUC better than 0.90. Moreover, after the training sample size goes beyond 3000, the performance gain with the training sample size will diminish. Their test dataset was composed of 500 chest radiographs of 500 patients. The COVID-19 and non-COVID-19 pneumonia ratio had the same proportion. A similar trend was observed in our results, where the number of training data for which the gradient disappears was approximately 1600, 400, 2200, and 1100 for AlexNet, Fine-tuned AlexNet, ResNet34, and Fine-tuned ResNet34, respectively (corresponding AUCs: 0.98, 0.99, 0.98, and 0.99, respectively). To generalize our results, additional experiments by using Fine-tuned AlexNet and ResNet34 were conducted to compare the model performance of using the datasets labeled as COVID-19 and Normal with that of using the datasets labeled as COVID-19 and non-COVID-19 pneumonia. Each dataset comprises 2500 CXR images that are randomly selected in the BrixIA and Chest X-ray 14 dataset. The dataset was split as 2000 images for training and validation, and 500 images for testing. Furthermore, training and validation data were selected as 15 different datasets. Fig. 10 shows the relation with AUC for COVID-19 and Normal and COVID-19 and non-COVID-19 pneumonia datasets. Similar trends are observed for both results. The AUC in the equilibrium state of the “pneumonia” dataset was almost the same as that of the “Normal” dataset, indicating good performance. Considering Ibrahim et al. [38], who achieved high performance using automatic detection AlexNet to classify CXR images of COVID-19 and non-COVID-19 viral pneumonia and COVID-19 pneumonia and healthy subjects, our results of binary classification for COVID-19 and Normal can be generalized to classify COVID-19 and non-COVID-19 pneumonia. Additional experiments also indicate that the main reason for the higher performance than Ran et al.’s result is not the difference in the target diseases but the collection of their dataset from multiple medical institutes.

Fig. 11 shows the relation of sensitivity, specificity, and training sample size using the nonlinear function obtained from Table 2. The sensitivity and the specificity of AlexNet outperformed ResNet34 with and without the fine-tuning method in the small number of CXR images. In addition, the difference in these performances tended to decrease as the number of training data increased. In other words, if the available data are limited, AlexNet is a more proper model to use compared to ResNet34. D’souza et al. [39] conducted structural analysis and optimization of convolutional neural networks with a small sample size because the number of samples in a dataset can be relatively limited in numerous real-world applications. They trained and tested each structure followed by layer dimension (layer width) optimization using small subsets of these datasets from entirely different data nature (calligraphic, photographic, and microscopic). Their result suggests that “deeper the better” is not always true for CNNs for small datasets, also clearly shows that as the depth increase there is an initial drop in the classification error, but the error soon rises sharply (calligraphic and microscopic). However, this may not always be true, as the microscopic dataset has no clear bias toward deeper or shallower networks. They concluded optimally performing network is largely determined by the data nature. Our result, which 8-layer AlexNet outperformed the
Our study has some limitations. First, our dataset consists of only COVID-19 and Normal, and the ratio was limited to the same proportion. In the real-world data, the training data are expected to include CXR images with many pathologies not limited to COVID-19 and Normal alone. The ratio of COVID-19 and other pathologies, including Normal, may also vary. Therefore, as a future work, we will conduct a study on the performance change when adding many pathologies to the training data and change the ratio in the training data. Second, our study focused on binary classification to classify COVID-19 and Normal. To achieve a more effective classification, it is desirable to validate multi-class classifications, such as COVID-19, other viral, bacteria, and Normal by using CXR images. Although Cohen et al. [40,41] made CXR images labeled as detailed Viral, Bacterial, Fungal, etc. available to the public, the number of images is very limited. These data are continuously being collected from public sources and through indirect collection from hospitals and physicians. Therefore, it is hoped that many of these data will become available in the future.

6. Conclusion

Appropriate performance changes must be predicted to manage the performance of medical devices using AI/ML. In this study, we performed binary classification to classify COVID-19 and Normal by using large datasets. In addition, we observed the performance changes (i.e., AUC, sensitivity, and specificity) with the change in the number of training data. In the medical field, comprehensive research is limited in the literature that uses large datasets, particularly on the relationship between performance and training data because building large datasets is costly and burdensome for professionals, and there are concerns about ethical and privacy issues. This paper serves as a fundamental insight for regulators, policy makers, researchers, and manufacturers on how to develop appropriate post-market performance change plans.

Declaration of competing interest

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] G. Ljèjm, T. Koo, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafourian, J.A.W. M. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88. Dec.
[2] Ryui Hamamoto, Kruhi Suvarna, Masayoshi Yamada, Kazuma Kobayashi, Norio Shinkai, Mototaka Miyake, Masamichi Takahashi, Shinichi Jinmai, Ryo Shionomyo, Akira Sakai, Ken Takasawa, Aino Bolotan, Kanto Shoun, Dozen AI, Hidenori Machino, Satoshi Takahashi, Ken Asada, Masaki Komatsu, Jun Sene, Syuyo Kano, Application of artificial intelligence technology in oncology: towards the establishment of precision medicine, Cancers 12 (12) (2020).
[3] Urs J. Mueebmatter, Paola Daniere, Kerstin N. Vokinger, Approval of Artificial Intelligence and Machine Learning-Based Medical Devices in the usa and Europe (2015–20): a Comparative Analysis, The Lancet Digital Health, 2021.
[4] U.S. Food and Drug Administration, Artificial Intelligence and Machine Learning (AI/ML) software as a Medical Device Action Plan, 2021. January.
[5] U.S. Food and Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/machine Learning (AI/ML)-based Software as a Medical Device (Samd) - Discussion Paper and Request for Feedback, 2019. April. Nat.
[6] World Health Organization, Use of Chest Imaging in Covid-19: a Rapid Advice Guide, 2020 June.
[7] American college of Radiology, Acr Recommendations for the Use of Chest Radiography and Computed Tomography (Ct) for Suspected Covid-19 Infection, 2020. March.
[8] Olga Russakovsky, Deng Jia, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[9] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging 9 (4) (2018) 611–629. April.
[10] Ezz El-Din Hemdan, Marwa A. Shouman, Mohamed Esmail Karar, Covids-net: A Framework of Deep Learning Classifiers to Diagnose Covid-19 in X-ray Images vol. 11055, 2020 arXiv:2003.
[11] Narin Ali, Ceren Kaya, Ziynet Pamuk, Automatic Detection of Coronavirus Disease (Covid-19) Using X-ray Images and Deep Convolutional Neural Networks, arXiv, 2003, p. 2020, 10849.
[12] Praabha Kumar Sathy, Santi Kumari Behera, Detection of Coronavirus Disease (Covid-19) Based on Deep Features, MDPI AG, 2020.
[13] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baboglu, O. Yildirim, U. Rajendra Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, Comput. Biol. Med. 121 (2020), 103792, 06.
[14] Linxia Wang, Alexander Wong, Covid-net: A Tailor-made Deep Convolutional Neural Network Design for Detection of Covid-19 Cases from Chest X-Ray Images, arXiv, 2003, 09871, 2020.
[15] F. Ucar, D. Korkmaz, COVIDiagnosisNet-Diagnosis: deep Bayes-SqueezeNet based diagnosis of coronavirus disease 2019 (COVID-19) from X-ray images, Med. Hypotheses 140 (2020) 109761. Jul.
[16] S.R. Nayak, D.R. Nayak, U. Sinha, V. Arora, R.B. Pachori, Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: a comprehensive study, Biomed. Signal Process Control 64 (2021) 102565. Feb.
[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. June.
[18] Alex Krizhevsky, Ilya Sutskever, Geoffrey H. E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.
[19] M.M. Rahman, C. Li, Y. Yao, F. Kuwata, M.A. Rahman, B. Wang, S. Qi, F. Kong, X. Zhu, X. Zhao, Identification of COVID-19 samples from chest X-ray images using deep learning: a comparison of transfer learning approaches, J. X Ray Sci. Technol. 28 (5) (2020) 821–839.
[20] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv:1409.1556.
[21] Tuan D. Pham, Classification of covid-19 chest x-rays with deep learning: new models or fine tuning? Health Inf. Sci. Syst. 9 (1) (2021) 1–11.
[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, Going Deeper with Convolutions, arXiv:1404.4824, 2014.
[23] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer, SqueezeNet: Alexnet-level Accuracy with 50× Fewer Parameters and ~0.5Mb Model Size, 2016 arXiv:1602.07360.
[24] L. Wynants, B. Van Calster, G.S. Collins, B.R. Riley, G. Heinze, E. Schuit, M.M. J. Bonvin, D.L. Dahlby, J.A.A. Dennen, T.P.A. Debray, V.M.T. de Jong, M. De Vos, P. Dhiman, M.C. Haller, M.O. Harbay, L. Henckaerts, P. Heus, M. Kammer, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, G.P. Martin, D.J. McLernon, C. L. Andaur Navarro, J.C. Reitsma, J.C. Sergeant, Ch. N. Skeot, I.J.M. Smits, K.I. E. Snell, M. Sperrin, R. Spijker, E.W. Steyerberg, T. Takada, I. Traulsl, S.M.I. van Kuijik, B. van Bussel, E.C.C. van der Horst, P.S. van Ruyven, J.Y. Verbakel, C. Wallisich, J. Wilkinson, R. Wolff, L. Hoof, K.G.M. Moons, M. van Smeden, Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ 369 (2020) m1328, 04.
[25] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica L. Aviles-Rivero, Christian Ettman, Cathal McGaughey, Lucian Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans, Nat. Mach. Intell. 3 (3) (2021) 199–217.
[26] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gervasi, M. Ravanelli, A. Borghesi, R. Maroldi, D. Farina, BS-Net: Learning COVID-19 diagnosis/detection products: insights from fda-approved products, Expert Rev. Curr. Med. Drugs 17 (9) (2020) 899–918. September.
[29] Matthew D. Zeiler, Rob Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.

[30] Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, How Transferable Are Features in Deep Neural Networks?, 2014 arXiv:1411.1792.

[31] Rajeev Kumar, Abhaya Indrayan, Receiver operating characteristic (roc) curve for medical researchers, Indian Pediatr. 48 (4) (2011) 277–287.

[32] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (3) (1983) 839–843. Sep.

[33] William J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.

[34] Yingtao Fang, Jianzhou Wang, Xiaomin Ou, Hongmei Ying, Chaosu Hu, Zhen Zhang, Wei gang Hu, The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients, Phys. Med. Biol. 66 (18) (2021) 185012.

[35] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadiiski, Mark A. Helvie, Caleb D. Richter, Kenny H. Cha, Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets, IEEE Trans. Med. Imag. 38 (3) (2018) 685–696.

[36] Nima Tajbakhsh, Jae Y. Shin, R Guru Suryakanth, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, Jianming Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imag. 35 (5) (2016) 1299–1312.

[37] Ran Zhang, Tie Xin, Zhihua Qi, Nicholas B. Bevins, Chengzhu Zhang, Dalton Griner, Thomas K. Song, Jeffrey D. Nadig, Mark L. Schiebler, John W. Garrett, et al., Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: value of artificial intelligence, Radiology 298 (2) (2021) E88–E97.

[38] Abdullahi Umar Ibrahim, Mehmet Ozsoz, Sertan Serte, Fadi Al-Turjman, Polycarp Shizawaliyi Yakoi, Pneumonia classification using deep learning from chest X-ray images during Covid-19, in: Cognitive Computation, 2021, pp. 1–13, pages.

[39] R.N. D’souza, P.Y. Huang, F.C. Yeh, Structural analysis and optimization of convolutional neural networks with a small sample size, Sci. Rep. 10 (1) (2020) 834, 01.

[40] Joseph Paul Cohen, Paul Morrison, Dao Lan, Karsten Roth, Tim Q. Duong, Marzyeh Ghassemi, Covid-19 Image Data Collection: Prospective Predictions Are the Future, arXiv 2006.11988, 2020.

[41] Joseph Paul Cohen, Paul Morrison, Lan Dao, Covid-19 Image Data Collection, 2020 arXiv 2003.11597.