Development of a Deep Learning Model for Chest X-Ray Screening

W.H. Hsu[1], F.J. Tsai[2], G. Zhang[3], C.K. Chang[4], P.H. Hsieh[5], S.N. Yang[5], S.S. Sun[6], KenYK Liao[7], Eddie TC Huang[5,7,8,*]

[1] Division of Pulmonary and Critical Care Medicine, China Medical University Hospital and China Medical University, Taichung, Taiwan
[2] Department of Medical Research, China Medical University Hospital, Taichung, Taiwan
[3] Department of Radiation Oncology, Moffitt Cancer Center, Tampa, FL, USA
[4] Department of Medical Imaging, Chang Bing Show Chwan Memorial Hospital, Changhua, Taiwan
[5] Department of Biomedical Imaging and Radiological Science, China Medical University, Taichung, Taiwan
[6] Department of Nuclear Medicine and PET Center, China Medical University Hospital, Taichung, Taiwan
[7] Artificial Intelligence Center for Medical Diagnosis, China Medical University Hospital, Taichung, Taiwan
[8] Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

Abstract— **Developed in recent years, deep neural network becomes the best method for rapid analysis of advanced features and automation in medical image analysis. As a second clinical opinion provided by artificial intelligence (AI), it can reduce the physician's workload and reduce misjudgment. This study collected 365,892 chest X-ray images and clinical diagnosis reports through retrospective analysis, and compared five different input image sizes and images that generated by clinical labeling pre-processing in the classification model building and testing. An AI trained chest X-ray abnormal interpretation model by using DesNet121 neural network gave a test accuracy of 0.875. Deep neural network shows the potential of accountable methods to help lung classifications for normal and abnormal screening in clinics.**

Keywords— **Deep learning, Chest X-Ray**

I. Introduction

The 2016 pneumonia mortality statistics ranked the top three causes of death in Taiwan. Besides left and right atrium, ventricular enlargement, aortic aneurysm, calcification or exfoliation couldn't be ignored in cardiomegaly diagnosis, which are frequently the disorders or diseases that cause heart failure (HF), often distinguished by severity level. Therefore understanding the complex interactions between the cardiopulmonary system is an indispensable part of the treatment of these patients [1]. Furthermore atherosclerosis is one of the major potential pathological processes leading to heart attack (coronary heart disease) and stroke (cerebrovascular disease) [2].

Because of its features and functions that reveal unforeseen pathological changes, non-invasive, low radiation doses, etc. [3], chest X-ray is the first choice for diagnosis pneumonia and other lung or heart diseases [4]. The most common diagnostic findings from chest X-ray images include pulmonary infiltration, abnormalities in catheter and heart size or contour [5]. Chest X-ray plays an important key role in clinical care and epidemiography research [6, 7]. However, detect and diagnose diseases in chest X-ray is a challenging mission that rely heavily on the clinical diagnostic experience of a professional radiology physician.

Beside X-ray in chest examination, another method is computed tomography (CT). CT also uses X-ray to penetrate the human body, and the signal data received by photodetectors being reconstructed to generate 3-dimensional (3D) images of the body. Today, CT data can provide accurate clinical diagnosis [8]. However, sometimes CT imaging needs to inject Iohexol to help highlight the disease site, which could cause serious allergy for some patients. Therefore, before the CT examination, patients must pass drug allergy test or risk assessment based on related medical history. Furthermore, patients who take CT examination are expected to receive much higher radiation dose than patients who take chest X-ray imaging, and the CT examination time is much longer. Because of these restrictions, chest X-ray is still the most used method in clinical examinations, including pneumonia and other lung and heart diseases [9]. Chest posterior-anterior (PA) X-ray imaging is the most often used chest X-ray photography technique.

The development of artificial intelligence (AI) had been frustrated in several generations until ImageNet classification competition in 2012 in which AlexNet top-5 error rate was 10% lower than previous year's champion [10]. Since then, convolutional neural network (CNN) has been received strong attention from researchers. AI is now widely applied, and its algorithms include deep learning, machine learning and natural language processing. In recent years' machine learning has been used in automatic detection, extraction and classification of tumors [11-14]. The newest algorithm improvement of deep learning and very large database can surpass professional personnel in medical image missions, including diabetic retinopathy detection, skin cancer classification, arrhythmia detection and hemorrhage identification [15-18].

Automated diagnosis of thoracic diseases gets highly attention in recent years. Professor Lakhani used CNN of deep learning, and developed automatic classification of

tuberculosis disease from chest X-ray [19]. Professor Huang used the algorithm to find features of CT images to detect and diagnose pulmonary nodules [20]. Moreover, some people used the data of Open-I study on the performance of various convolutional structures on difference of abnormal diseases [21]. Subsequently, Professor Wang released ChestX-ray-14, and it had further development in thorax diseases diagnosis. ChestX-ray-14 had a much larger amount of data than previous data of the same type, and they also benchmarked different CNN frame pre-trained on ImageNet [22]. Subsequently, based on this data set, some people developed a method, CheXNet, that was better than the previous algorithm in diagnosis of 14 kinds of chest diseases [23, 24]. However, the best accuracy of all the results was only about 80% so far, and obviously there is still room for improvement.

Traditional computer-aided methods use algorithms to assist disease diagnosis. Deep learning methods should be doing better. Professor Li compared deep learning and feature-based statistical learning in breast density assessments, which demonstrated that deep learning was better than feature-based statistical learning [25]. Anticipating the application of deep learning to the detection of breast density in the future, and applying the model to the prediction of breast cancer risk, it is expected once again to enhance the feasibility of using AI in the medical fields. Although chest X-ray and computed tomography are the major tools of diagnosis in thorax, according to the estimation of World Health Organization (WHO), there are two-third people couldn't get radiodiagnostic resources in the world [26]. Even through there are imaging devices available, there is a lack of experts who can interpret X-ray images, resulting in increased mortality from treatable diseases [27]. With expert-level automation, this research aimed at the evaluation of the effectiveness and feasibility of deep learning neural network applications in chest X-ray interpretation through the combination of AI and medical cross-domain. The technology developed in this research project was hoped to improve health care services in the future, and eventually provide diagnostic and treatment help in areas where the number of professional radiologists is limited, while giving the region the opportunity to learn medical imaging expertise.

II. Materials and methods

Data collection: This research retrospectively collected data from the Hospital of China Medical University, including a total of 365,892 subjects in two years between 1/1/2017 and 12/31/2018, for the model training and test. And a total of 1,883 data in January 2019 were used for the final model evaluation. The data for each case included one chest X-ray image and a corresponding radiology report. The original image was in Digital Imaging and Communications in Medicine (DICOM) format. There were no gender and age restrictions in the data collection. This research was approved by the Institutional Review Board (IRB: CMUH106-REC1-092).

Exploratory data analysis (EDA): The 365,892 radiology reports of chest X-ray were initially filtered through text mining to select key words include PA View or Chest PA. There were 80,246 chest X-ray and reports with PA View or Chest PA. At the same time, in the "Protocol Name" of Dicom Header in the image, the chest X-ray image of the Posterior to Anterior View (PA View) was selected by "Chest PA", and the unsuitable image data, such as AP View and KUB, were excluded. Comparing reports and images, a total of 44,430 chest X-ray were selected for the study, as shown in Table 1.

Table 1 Data selection

Data File Name	Number of reports	Report included ”PA View” or ”Chest PA”	Corresponding image with Protocol Name included ”PA View”
106_1	74549	20802	12638
106_2	81697	20762	6310
106_3	29758	4706	564
107_1	63927	11125	8479
107_2	30210	2843	1623
107_3	85751	20008	14816
Total	365892	80246	44430

Subsequently, the normal and abnormal X-ray images were classified by the radiologist. The normal data judgement was based on the report. There must be no additional findings. For example, like metal necklace, artificial implant and old fractures etc. would be classified as abnormal data. Finally, 9,322 chest X-ray data were classified as normal and 1,935 chest X-ray data as abnormal. These 11,257 data sets were used as the first preliminary test for mini database. Through Python language, the most used Scilit-Learn modules of KFold clustering was applied to divide the data into 10 equal parts with the same proportion. Among it, 7,867 were selected as training data sets, 2,257 for validation, and 1,132 as reference data for testing the merits of the model. Subsequently, the 1,883 cases collected in January in 2019 were filtered using the same analysis, resulted in the selection of 1,721 evaluation data for the final output models. This set of data was not applied in model training, but only used in the testing of the final model. The study flow chart is shown in Figure 1.

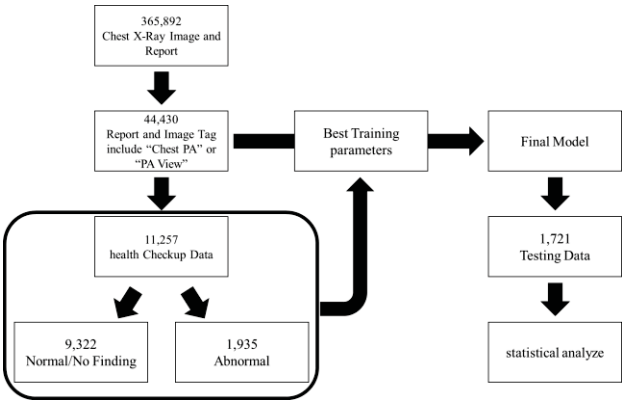

Fig. 1 Research process flow chart

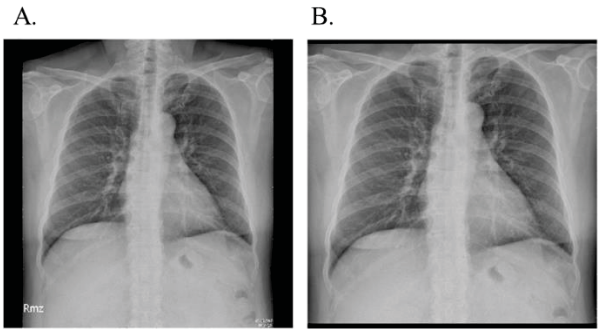

Fig. 2 A chest X-ray image after elimination of Annotation Marker. A. The original image, B. Image with cropping

Image Preprocessing: The Hospital of China Medical University has multiple institutions, including LIFI medical building, Children Hospital, rehabilitation medical building, MEIDE medical building, cancer center building and critical care center building. As a result, different brands of X-ray machines are used in the department of radiology in different buildings. The X-ray image machines are listed in Table 2. Because of the different equipment, different X-ray image sizes were common. Even with the same machine, different X-ray image sizes were also common because of different operations. In this study, all X-ray images were readjusted to the size of 224×224 or 299×299 pixels following the advice of Neural Networks model. In the same time, through DICOM header information, the contrast and brightness correction was also performed. The grey scale in X-ray images was converted to chromatic colors for the software analysis.

Table 2 X-ray machine brands in the hospital

Brand	Model
TOSHIBA	KXO-32R
TOSHIBA	KXO-50R
TOSHIBA	MRAD-A50S
TOSHIBA	MRAD-A80S
SHIMADZU	UD150L-RII
SHIMADZU	UD1506-RII

In addition to the purpose of image size standardization for the input, in the preliminary analysis, the annotation markers of the chest X-ray images were eliminated in the image resizing process. An annotation marker eliminated image is shown in Figure 2.

Deep Learning Model: Convolutional neural networks are the result by learning how the human brain works. The nerve cells exist in a human brain. The nerve cell uses synapsis to connect and receive external signal and pass to next neurons. Every neuron has different ability of conversion and with the information transfer and collection, human brain has the ability of thinking and judgement. Because of the rise of deep learning in recent years, CNN offered a lot of help in image recognition. CNN can automatically learn and identify features. It is suitable for 2D images. The most important feature which makes people impressed is the capability of generalization of another image identification.

CNN has variety types and different functions. This research was based on the multiple types of CNN deep learning system, and used the published chest X-ray disease model by Rajpurkar et al. developed using DenseNet121 as a reference. The original data were pre-processed using different methods and verified by the corresponding radiologists before the input of imaging data. The clinical reports were used as a basis for learning and training. Finally, the advantages and disadvantages of each model resulted were compared, and an optimal chest X-ray abnormality interpretation model was finalized. The overall training process is shown in Figure 3.

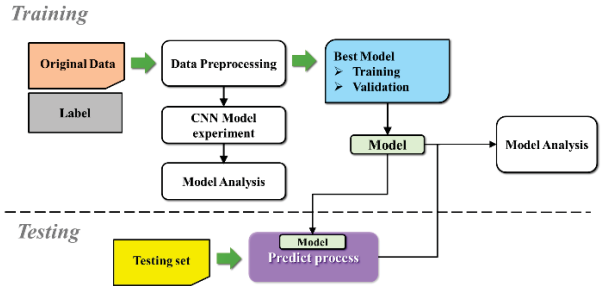

Fig. 3 Overall training process

Statistical Analyze: There were several criteria used for evaluating the pros and cons of the models: accuracy, area under curve (AUC), precision, recall (sensitivity), and F1

score. The confusion matrices were applied to calculate the actual true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values of the models in the test data set.

III. Results

First, 11,257 health check data were used for analysis. Under the same condition of neural network structure and hyper parameters, the most suitable model training method was evaluated by changing the different sizes of the image input and whether to perform an annotation marker removal image pre-processing process. For the 1,132 test set data, the input image was re-sized to 224×224, 224×224 with image pre-processing, 448×448, 512×512 and 672×672, respectively. Table 3 shows the results of accuracy, AUC, F1 Score, precision and sensitivity (Recall).

Table 3 Training result in different size of X-ray and image cropping

	224×224	224×224 with image cropping	448×448	512×512	672×672
Accuracy	0.712	0.602	0.635	0.570	0.649
AUC	0.711	0.681	0.688	0.700	0.708
F1 Score	0.402	0.351	0.375	0.384	0.404
Precision	0.315	0.245	0.267	0.256	0.287
Sensitivity/Recall	0.558	0.619	0.629	0.771	0.685

Confusion matrices and ROC curves on the test data for the five models are shown in Figures 4 and 5 respectively.

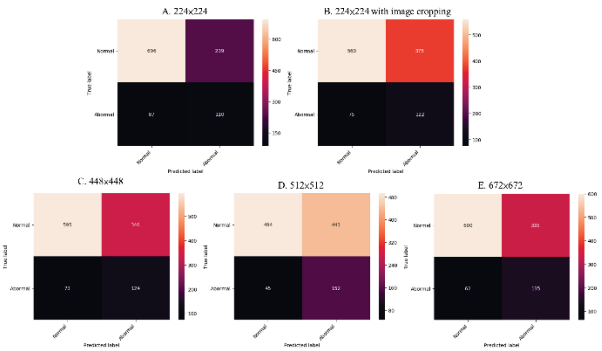

Fig. 4 Confusion matrix with different image size and image cropping

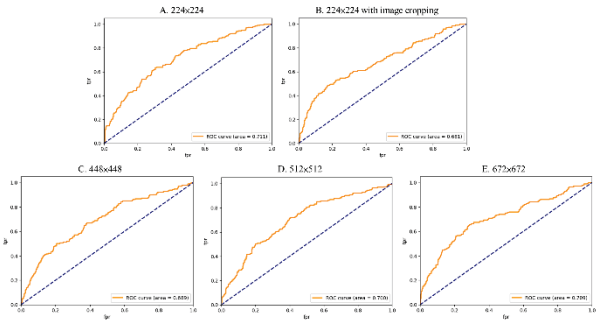

Fig. 5 ROC curves with different image sizes and image cropping

After the preliminary analysis, the 44,430 available data were screened and applied in the further training. The obtained final model was evaluated using the subsequent collection of 1,721 available data. The confusion matrix and ROC curve results of the evaluation are shown in Figure 6. The overall test accuracy rate reached 0.875, the AUC was 0.876, and the F1 Score, precision and sensitivity were 0.666, 0.738 and 0.606, respectively.

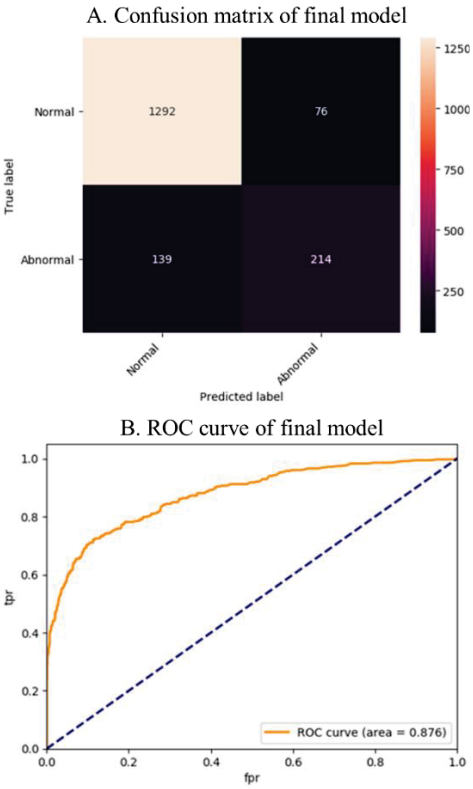

Fig. 6 Confusion matrix and ROC curve of final model using the evaluation data set

IV. Discussion

Comparison: Among the five models established using the 11,257 data of health check, the neural network architecture used was DenseNet121. The research found that the accuracy of model training results by changing the input image size was 71% optimal at image size of 224 × 224, and the training results at other image sizes did not improve. In addition, in the comparison of AUC, F1 Score, precision and sensitivity (Recall), all the results were better for the image size of 224 × 224 than the other input image sizes, with an exception of sensitivity. It shows that maintaining a training pattern that is closer to the original image size does not help the overall model optimization. If focused on AUC and the F1 Score, which takes into account both accuracy and sensitivity, the model with 224 × 224

image size still had higher scores. Further test based on this input image size was to remove the annotation markers in the input images. The results from Table 3 show that the model trained after the addition of the cutting process had a significant deterioration overall. The position of the neural network model reference was also visualized through the Gradient Class Activation Mapping (Grad-CAM) method. Figure 7 shows a case of normal chest X-ray images randomly selected from the test set. Image A is the original chest X-ray image, B is the color-highlighted image of Grad-CAM after image cropping and C is the Grad-CAM image without image cropping. In the Grad-CAM image, the closer the highlight color is to red, the higher the judgment value of the part affecting the neural network, which is the basis for the model to be "seen" by normal or abnormal interpretation. After the image cutting process, the test image displayed that the model focused on the black bar generated close to the abdominal cavity (Figure 7B). Therefore, the image characteristics of the processed chest X-ray were not correctly learned and thus misleading. The Grad-CAM results produced by the model trained without images cropping had strong visual representations in the bones, lungs, mediastinum, etc., which was more logically compatible with clinical interpretation (Figure 7C). Based on the observations of the two models, it was inferred that the black bars generated after image cropping might be one of the main causes of interpretation errors. It could be improved if the black bar was removed after the cropping. However, in this case, the image size must be resizing back to 224×224, which would introduce the image deformation. In addition, we also found that the image cutting method often cut off part of the lungs. As shown in the images of the red boxes in Figure 8, part of the apex of the lungs in the images was removed by this method. This is not allowed in clinical chest X-ray imaging standards. So we believed that this approach could make the neural network model erroneously learning. In summary, in the subsequent further model training, the pre-processing method of image cropping was abandoned.

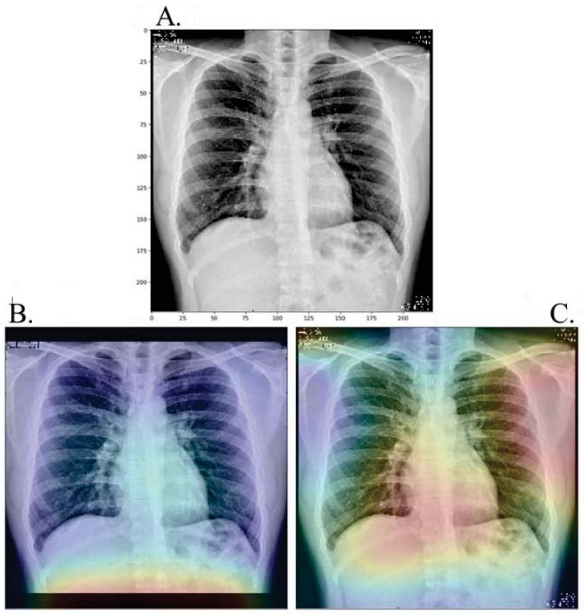

Fig. 7 An example of normal chest X-ray images. A. The original image, B. Grad-Cam image with cropping, C. Grad-Cam image without cropping

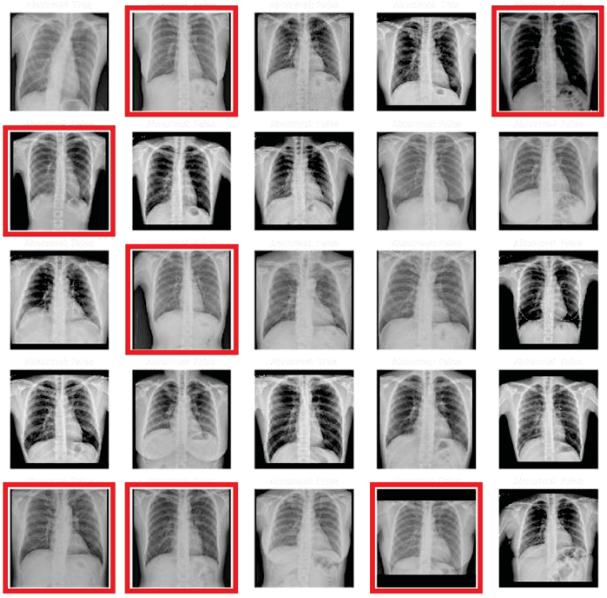

Fig. 8 Random sampling of chest X-ray images after image cropping. Part of the apex of the lungs were removed after cropping for the ones in red box

After preliminary training from the preliminary model, the decision was made to input image size 224x224 and not to crop image for final model training using all the subsequent training data. With the 44,430 chest X-ray data trained final model, the testing result using a month's chest X-ray interpretations collected in 2019 was 0.875 in accuracy, 0.876 in AUC. Overfitting was not observed from the ROC curve. Usually, the accuracy of higher than 0.8 is considered outstanding result. However, the F1 Score and

precision were 0.666 and 0.738 respectively. These values are lower than the conventionally recognized normal standard. From this point of view, this model still has some room for optimization and improvement.

Limitations of the Research: In this study, the most rigorous definition was applied in the normal image classification. When there was an obvious conflict between the clinical report and X-ray image: no finding or normal in the report but abnormal chest X-ray image, patient history was reviewed, as in the former literature, it has been proved that history of patients would affect the accuracy of radiologist interpret chest X-ray [28, 29] . Based on the text mining in patient's history, this kind of radiologist's reports was often found not exactly correct. It couldn't exclude the possibility of mistype report, or the physician determined that the symptoms were mild and thus gave a normal report. To correct these reports, data collection and the threshold of the need of the related cross discipline knowledge was really high. And it would be a difficult project to develop. The related methods of such corrections were not found in former literature. Therefore, those cases with questionable reports were excluded in this study.

V. CONCLUSIONS

In this paper, a chest X-ray assisted classification model using deep neural network is presented. This model is based on whether or not any abnormalities are mentioned in the radiology report. In addition, the performance of the DenseNet121 model under different input sizes and image cropping of the chest X-ray image was tested in this study. Finally, our own deep neural network classification model, which had a good performance in interpreting the abnormality of the chest X-ray film, was developed. This study further proved the feasibility of deep learning in the field of medical imaging classification.

ACKNOWLEDGMENT

This study was financially supported by China Medical University Hospital (CMU103-BC-1) and Chang Bing Show Chwan Memorial Hospital (BRD108006). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

53. Olson, T. and Johnson, B. (2011). Influence of cardiomegaly on disordered breathing during exercise in chronic heart failure. European Journal of Heart Failure, 13(3), pp.311-318
54. Mendis, S., Puska, P., Norrving, B., & World Health Organization. (2011). Global atlas on cardiovascular disease prevention and control. Geneva: World Health Organization
55. Campadelli, P., & Casiraghi, E. (2005, August). Lung field segmentation in digital postero-anterior chest radiographs. In International Conference on Pattern Recognition and Image Analysis (pp. 736-745). Springer, Berlin, Heidelberg
56. World Health Organization. (2001). Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children (No. WHO/V&B/01.35). Geneva: World Health Organization
57. van Ginneken, B., Hogeweg, L. and Prokop, M. (2009). Computer-aided diagnosis in chest radiography: Beyond nodules. European Journal of Radiology, 72(2), pp.226-230
58. Franquet, T. (2001). Imaging of pneumonia: trends and algorithms. European Respiratory Journal, 18(1), pp.196-208
59. Cherian, T., Mulholland, E. K., Carlin, J. B., Ostensen, H., Amin, R., Campo, M. D., ... & O'Brien, K. L. (2005). Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. Bulletin of the World Health Organization, 83, pp.353-359
60. Sharma, S., Maycher, B., & Eschun, G. (2007). Radiological imaging in pneumonia: recent innovations. Current opinion in pulmonary medicine, 13(3), pp.159-169
61. Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., & Rosenow III, E. C. (2012). Interpretation of plain chest roentgenogram. Chest, 141(2), pp.545-558
62. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems pp. 1097-1105
63. Bengio, Y. (2009). Learning deep architectures for AI. Foundations and trends® in Machine Learning, 2(1), pp.1-127
64. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), pp.1798-1828
65. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117
66. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), pp.436
67. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Kim, R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), pp.2402-2410
68. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), pp.115
69. Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836
70. Grewal, M., Srivastava, M. M., Kumar, P., & Varadarajan, S. (2018, April). Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 281-284. IEEE
71. Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 284(2), pp.574-582
72. Huang, P., Park, S., Yan, R., Lee, J., Chu, L. C., Lin, C. T., ... & Hales, R. (2017). Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. Radiology, 286(1), pp286-2951
73. Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... & McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2), pp304-310
74. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097-21061
75. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., & Lyman, K. (2017). Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.105011
76. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Lungren, M. P. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.052251

77. Li, S., Wei, J., Chan, H. P., Helvie, M. A., Roubidoux, M. A., Lu, Y., ... & Samala, R. K. (2018). Computer-aided assessment of breast density: comparison of supervised deep learning and feature-based statistical learning. Physics in Medicine & Biology, 63(2), pp.250051
78. Mollura, D. J., Azene, E. M., Starikovsky, A., Thelwell, A., Iosifescu, S., Kimble, C., ... & Johnson, B. (2010). White paper report of the RAD-AID Conference on International Radiology for Developing Countries: identifying challenges, opportunities, and strategies for imaging services in the developing world. Journal of the American College of Radiology, 7(7), pp.495-5001
79. Kesselman, A., Soroosh, G., Mollura, D. J., Abbey-Mensah, G., Borgstede, J., Bulas, D., ... & Fuller, L. (2016). 2015 RAD-AID conference on international radiology for developing countries: the evolving global radiology landscape. Journal of the American College of Radiology, 13(9), pp.1139-1144
80. Berbaum, K., Franken, J. E., & Smith, W. L. (1985). The effect of comparison films upon resident interpretation of pediatric chest radiographs. Investigative radiology, 20(2), pp.124-128
81. Potchen, E. J., Gard, J. W., Lazar, P., Lahaie, P., & Andary, M. (1979, January). Effect of clinical history data on chest film interpretation-direction or distraction. In Investigative Radiology (Vol. 14, No. 5, pp. 404-404). 227 EAST WASHINGTON SQ, PHILADELPHIA, PA 19106: LIPPINCOTT-RAVEN PUBL

Contacts of the corresponding author:

Author: Tzung-Chi Huang, Ph.D.
Institute: Artificial Intelligence Center for Medical Diagnosis, China Medical University Hospital, 2, Yude Road, 40447, Taichung Taiwan
Email: tzungchi.huang@mail.cmu.edu.tw