J MEDICINE 2023; 24: 112-118

Development and Validation of a Machine Learning System for Analysis and Radiological Diagnosis of Digital Chest X-ray Images

Mehrunnissa Khanom[1], Aseef Iqbal[2], Afroza Hoque[3], Fatiha Tasmin Jeenia[4], Muslim Uddin[5]

Abstract

Introduction: *Medicine is identified as one of the most promising field of application for Artificial Intelligence (AI). The current research presents development and validation of a machine learning system to diagnose chest x-ray images with higher accuracy.*

Methodology: *It was a multi-cantered, experimental study conducted from 01 July, 2021 to 30 June, 2022. The experiment was a two-step process; in the first step, a machine learning system (MLS) was developed through training, testing and tuning a specialised computer hardware & software utilising 5600 chest X ray images from NIH (National Institute of Health) chest X ray dataset. In the second step, 500 unseen chest X ray images from study centres were allowed to be diagnosed by the machine learning system and results were compared with expert opinions.*

Result: *After the system was developed, validation was done on 3 different variations of Deep Residual Network and tested for their accuracy in classifying the findings. Using ResNet50V2, an average accuracy of 84.37%% was achieved. With case-specific variation, highest accuracy was 94.84%, highest specificity was 97.23% and highest sensitivity was 88.25%.*

Conclusion*: With utilisation of this machine learning system, a faster radiological diagnosis of huge number of X ray images will become possible using only a small computer. Dependency on manpower, logistic support as well as rate of human-made errors can be minimised. However, this machine is never meant to replace an expert human opinion and it can never think beyond the box.*

Keywords: *Machine learning system, radiological diagnosis, digital chest X-ray.*

DOI: https://doi.org/10.3329/jom.v24i2.67273

Received: *25.05.2023;* ***Accepted:*** *10.06.2023*

Introduction

Artificial Intelligence (AI), since its inception traced back to 1956, has been an emerging field of technology that attempts to build intelligent decision systems from the information available around. Although much progress has been made in past decade, AI has mostly been capsuled among limited researchers from computer science due to inconsistent definition and evolving scope of area it encompasses. With the advent of modern equipment capable of producing digitized data and technological breakthroughs in computational infrastructures, AI has re-emerged into the consciousness of scientific and research communities.

Medicine is identified as one of the most promising field of application for AI – varying from Clinical Decision Support Systems [1–3], Rule-based diagnosis [4–6] and very recently to Machine learning techniques for clinical image classification [7–11] and automated interpretation of ECGs [12–14].

Recent AI research has leveraged machine learning methods with the availability of huge amount of data being generated every day and is gradually changing the landscape of healthcare and medical practices by identifying patterns from these raw data. One dominant application relevant to healthcare and medicine is to take a large number of 'training'

1. Associate Professor Department of Medicine, Chattagram International Medical College
2. Associate Professor, Department of Computer Science and Engineering, School of Science & Engineering, Chittagong Independent University
3. Assistant Professor, Department of Medical Education Unit, Chattagram International Medical College
4. Associate Professor, Department of Pharmacology, Chattagram International Medical College
5. Professor, Department of Paediatrics, Chattagram International Medical College

Author of Correspondence: Dr. Mehrunnissa Khanom, Associate Professor Department of Medicine, Chattagram International Medical College, Email: drmehrun.k@gmail.com Phone: 01713109200

cases as input (photographs of fundus, x-ray images of different body parts, etc.) and label them into predetermined classes by analyzing the patterns of these data. The algorithm learns the pattern and can take a 'new' data of similar case and categorize it accordingly. These algorithms are designed to identify the optimal parameters in the models to minimize the deviations between their predictions for the training cases and the observed outcomes in these cases, with the expectation that the identified associations are generalizable to cases not included in the training dataset. A special case for such machine learning algorithm is "Deep Learning' which involves training an artificial neural network with many layers large-scale annotated image dataset[15]. With the help of modern machine-learning methods, a number of radiology applications, such as diagnosis of pulmonary tuberculosis and other common lung diseases with chest radiography[16], breast-mass identification using mammography scans [17], detection and management of lung nodules [18], etc. have reached expert level accuracies.

The current paper presents an experimental study to develop a machine learning system employing deep learning techniques to automatically annotate a chest X-ray image into one of the pre-trained disease classes utilizing publicly available radiology dataset and compares the machine learning with expert opinion.

Materials and methods

It was an experimental study conducted from 01 July, 2021 to 30 June, 2022 at four private centers of Chattogram city including one non-government medical college. In this study, a specialized computer with specialized hardware (Quad-core CPU & dedicated high performance GPU) and software (Tensorflow, Keras, OpenCV) was used as research instrument.

There were two basic steps of experimentation: development phase and validation phase.

In the development phase, a system architecture was developed (Figure 1), which pre-processed, trained and tuned the machine learning system with chest X-ray images from NIH dataset.[19] After machine learning system became ready to read the unseen data, it went through the validation phase. It was then provided with unseen samples randomly selected from four study centers. The machine recognized and classified the unseen samples into disease or diagnosis classification according to the dataset; the result from machine learning was verified for accuracy by matching with result from expert opinion. Chest X-ray images collected from the study centers were used as ground truth data, which were labeled by the experts of radiologists in the study centers, having at least 5 years of experience in the respective field. Statistical analysis was performed by Scikit-learn library (built-in statistical software of machine learning system).

The sample size for development phase was 5600, these were digital chest X-ray images selected from NIH (National Institute of Health) digital chest X-ray dataset by simple random technique. Out of these 5600 images, 60% were used for training and 40% were used for testing & tuning the system. Sample size for validation of developed machine learning system was 500, these were digital chest X-ray images of adult (≥18 years) selected from four study centres by stratified random sampling. Inclusion criteria for the samples for validation phase were availability of complete record (Id no, expert opinion), within last three months of commencement of study and image with a specification criteria of DICOM (Digital Imaging & Communications in Medicine) format. Images with incomplete record, poor technical quality and without specification criteria were excluded from the study. The ethical clearance and permission letters were taken from all participating centres, the confidentiality of patients was strictly respected during handling of data; only anonymous data was utilized.

NIH dataset: National Institutes of Health (NIH) is a collaborating research agency constituting 27 different medical institutions and centers and operates under the U.S. Department of Health and Human Services. NIH has developed a dataset comprising 112,120 chest X-ray images classifying fourteen different thoracic pathologies. These diseases are: atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodules, pleural thickening, pneumonia and pneumothorax. Each X-ray image is of 1024x1024 resolution and has associated meta data containing image index, view position, finding labels (expert opinion), patient ID, age, gender along with the size of the image file.

The NIH dataset used were collected from Kaggle which is made available by the NIH for public use. This dataset is developed following the dataset construction method by NIH, which in turn, is a reliable source of valid data.

Machine Learning Process:

Developing any machine leaning system involves partitioning the dataset into 2 parts: training and testing.

Training data: A major part of the samples (x-ray images) of the dataset is used to train a machine learning system to generate a model by fitting the parameters to classifier.

Test data: The final part of dataset with yet unseen samples that might be used to test the performance of final model for unbiased evaluation. It is based on the performance of this set of data where we identify the accuracy or efficiency of a machine learning system for a specific case or scenario.

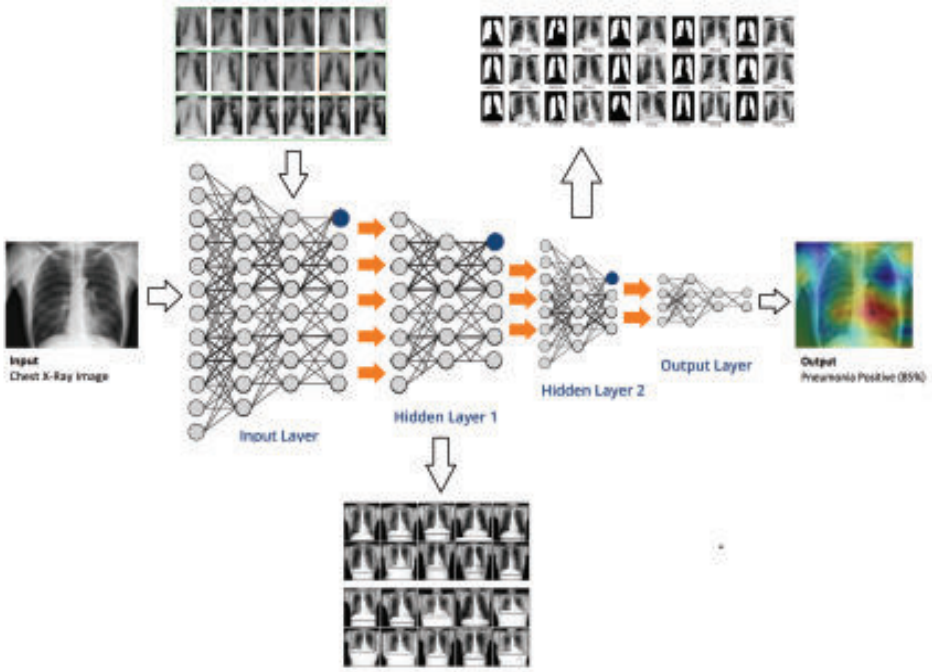
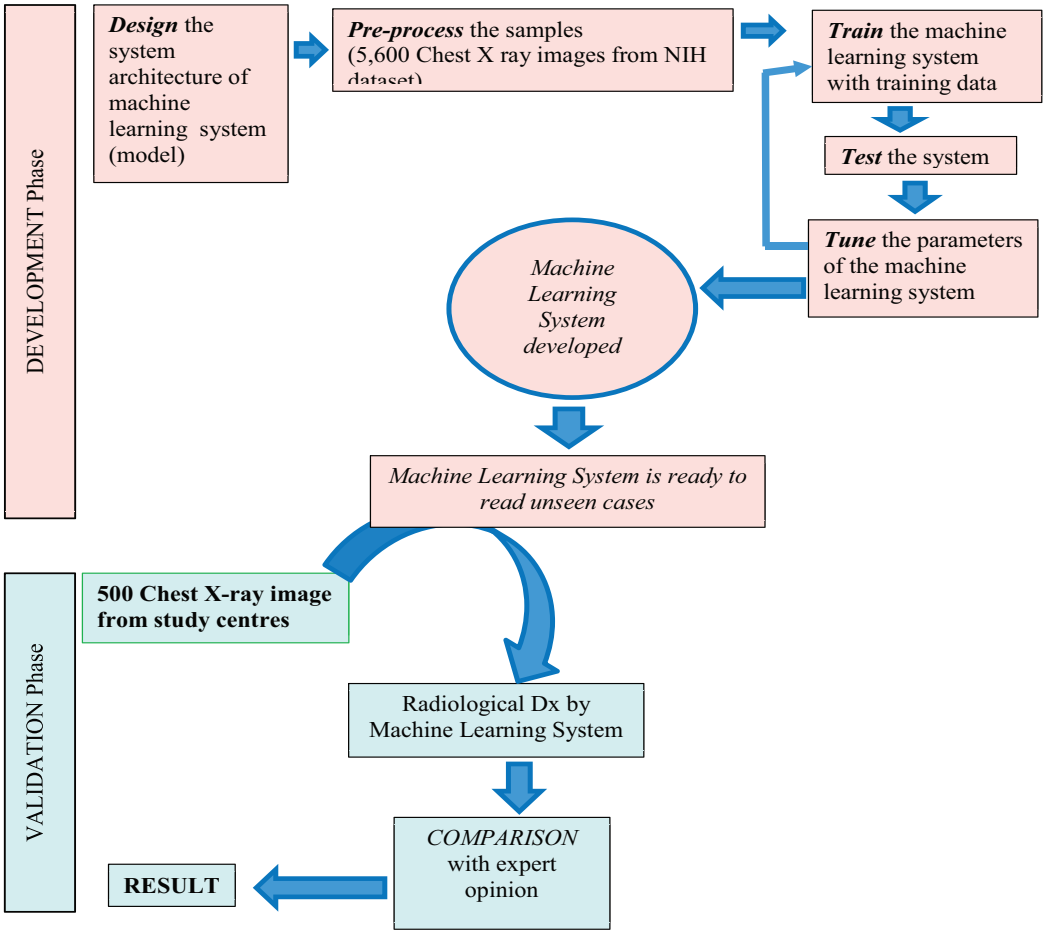


Figure 1: *Proposed system architecture of the machine learning system*

Flow chart:



Results

Data Samples for machine learning

For this machine learning project, a total of 5600 sample x-ray images were randomly picked from NIH dataset. The system was trained using 60% of the total samples available in the dataset (chosen by the computer at random to avoid any biasness), then remaining 40% data samples were used to see how the system was performing by comparing machine generated result with the result in dataset and tuning the parameter values if necessary.

Pre-processing of the data:

The chest x-ray images in NIH dataset were collected from different test centers and had variation in orientation, brightness, contrast, size, etc. which might adversely affect the machine learning model. So, it was necessary that the input data were all normalized before being used for training a machine learning system. For this study, the following preprocessing steps were performed on NIH dataset images before training:

Step 1: Resizing images – At this step the resolution of input images of NIH dataset was reduced to 224x224 from 1024x1024. This was done to reduce the computational overhead without compromising on classification accuracy.

Step 2: Data Augmentation – A common technique to improve the robustness of a deep learning based system is to train the system with some augmented data generated from training samples. This helps the machine learning system to perform with expected accuracy even if the input sample is not exactly same as the training samples in terms of scaling, cropping, flipping or rotation.

Step 3: Checking for data imbalance – Ideally a machine learning system would have best training if it is trained with equal number of samples from each class to avoid biasness. However, almost in all cases, the datasets developed from field data would have skewed class proportions – the NIH Chest X-ray dataset is also no exception. Data imbalance can be seen from the figure 2, making up majority classes (higher proportion of training data) and minority classes (lower proportion of training data). The class 'No_findings' represents the group presented with no abnormality is chest x-ray.

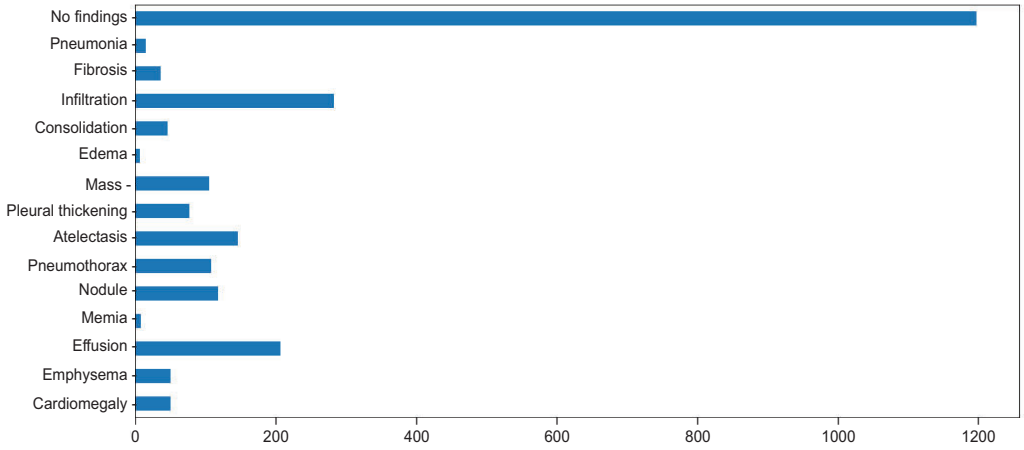


Figure 2: *Data imbalance among the classes (disease categories)*

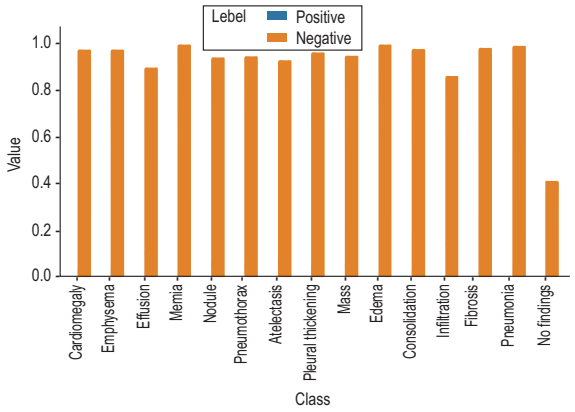


Figure 3: *Bar chart showing frequency of positive and negative images of each disease class.*

A widely adopted technique of dealing with data imbalance was to multiply the frequency of positive and negative images of each class with a weight value obtained from the frequency of opposite label. In our machine learning development, this operation is performed as demonstrated in the code segment below written in Python Language:

```
pos_weights = freq_neg
neg_weights = freq_pos
pos_contribution = freq_pos * pos_weights
neg_contribution = freq_neg * neg_weights
```

The chart in figure 4 showed that the actual impact during the training period would be the same for negative and positive cases across all diseases. At the end of this stage, the data was ready to train the machine learning system.

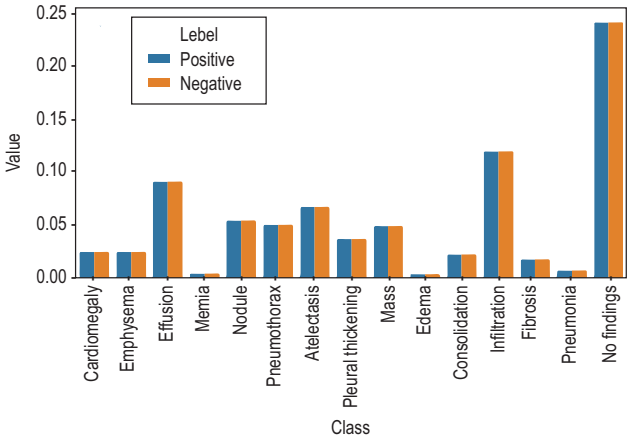


Figure 4: *Bar chart showing frequency of positive and negative images after dealing with data imbalance*

Training the Machine learning system

For this project, **ResNet50V2,** known for their accuracy in classifying image data was adopted to develop the machine learning models.

Epochs – Similar to the human learning process, a deep neural network based machine learning system tends to improve performance if it is trained with the entire set of training data for a number of times. This process of training a machine learning system with entire training dataset is called epochs. A machine learning system may take a number of epochs, which depends on the problem at hand and type of input data, to reach of optimum accuracy.

Batch size – Due to huge computational overhead (Processing, memory, storage), the training dataset is split into batches of smaller data before being fed into the machine learning system. For this research project, the training data of 3360 images were split into batches of 56 images, and all the batches of images were fed to the system 30 times (epochs). If the curve of training accuracy is not stabilized by the end of 30 epochs, we'd need to increase the number of epochs gradually until the curve is not rising anymore.

Testing the machine learning system

After the training phase, the system has been tested and tuned for accuracy using the remaining 2240 images from the dataset that the system has not encountered before. The same batch size of 56 were used and the testing were also performed after each of 30 epochs of training.

Validation of the machine learning system

Finally, the developed system has been evaluated for accuracy in classifying the chest x-ray images. At this stage, the system was ready to take unseen chest x-ray images randomly from study center as per inclusion criteria and make prediction (i.e. classification) of the input sample into one of the recognized classes that returns with highest probability.

A total of 500 digital x-ray samples were collected from the test center for this project. The x-rays collected were in DICOM format which is standard output from any digital x-ray machine. These DICOM files were converted to image files manually and resized to 224x224 pixels before they were put through the systems for validation.

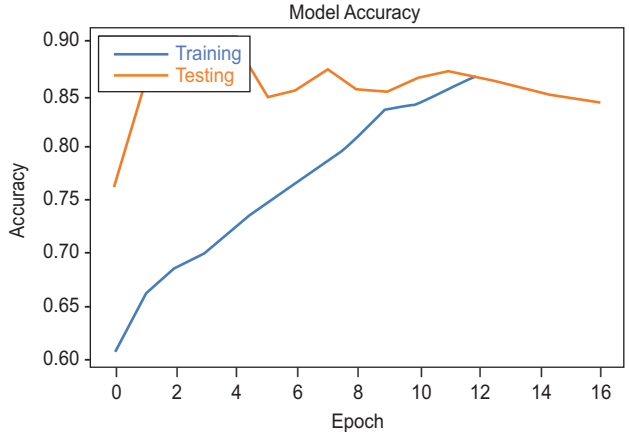


Figure 5: *Relationship between accuracy obtained with training and testing during development phase*

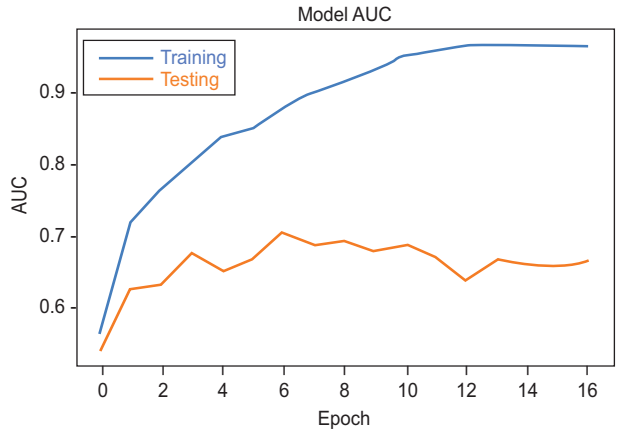


Figure 6: *Relationship between the area under the curve (AUC) obtained with training and testing during development phase*

Figure 5 shows the relationship between accuracy obtained with training and testing data after each epoch. Training accuracy started improving from zero, then came very close to 1; testing accuracy started above 76% and it kept improving in a range. For multi-labelled data, AUC is a better indicator of performance of a machine learning system than accuracy. Figure 6 shows relationship between the area under the curve (AUC) obtained with training and testing data after each epoch. During training, the receiver operating characteristics (ROC) reached very close to 1 at the end of 16th cycle. At the same time, the ROC of testing remained close to 0.7, at this stage performance 70% of predicted, starting from zero.

Table I: *Accuracy, sensitivity & specificity of the developed machine learning system*

Class	Accuracy %	Positive Predictive Value	Negative Predictive Value	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood Ratio
Cardiomegaly	84.21875	0.072164948	0.979742173	38.88889	85.53055	2.687654321	0.71449457
Emphysema	94.84375	0.166666667	0.978896104	23.52941	96.78973	7.329411765	0.790069262
Effusion	79.6875	0.264900662	0.961145194	67.79661	80.89501	3.548633379	0.398088713
Hernia	93.28125	0	0.998327759	0	93.42723	0	1.070351759
Nodule	83.90625	0.0625	0.925347222	8.510638	89.88196	0.841134752	1.017883518
Pneumothorax	86.25	0.1625	0.9625	38.23529	88.94389	3.458296752	0.694423224
Atelectasis	74.6875	0.188571429	0.956989247	62.26415	75.8092	2.573877226	0.497774009
Pleural_Thickening	93.59375	0.055555556	0.961414791	4	97.23577	1.447058824	0.98729097
Mass	91.875	0.25	0.944805195	15	97	5	0.87628866
Edema	89.84375	0	0.998263889	0	89.98435	0	1.111304348
Consolidation	93.4375	0.090909091	0.980230643	20	95.2	4.166666667	0.840336134
Infiltration	58.4375	0.164794007	0.884718499	50.57471	59.6745	1.254162157	0.828247997
Fibrosis	87.34375	0.042857143	0.975438596	17.64706	89.24559	1.640913082	0.922767668
Pneumonia	88.28125	0	0.99122807	0	88.97638	0	1.123893805
No_Findings	65.78125	0.647294589	0.695035461	88.25137	35.76642	1.373913313	0.328482213

In the validation phase, the average accuracy of the system using ResNet50V2 was 84.364583333%. Table I shows the case-specific accuracy, positive predictive value, negative predictive value, sensitivity, specificity, positive likelihood ratio (LR) and negative likelihood ratio of the developed machine learning system. A likelihood ratio of 1.0 indicates that there is no difference in the probability of the particular test result (positive result for LR+ and negative result for LR") between those with and without the disease. A likelihood ratio >1.0 indicates that the particular test result is more likely to occur in those with disease than in those without disease, whereas a likelihood ratio <1.0 indicates that the particular test result is less likely to occur in those with disease than those without disease. As LRs move farther away from the value of 1.0, the strength of their association with the presence or absence of disease increases.

Thus, tests with very high LR+ and very low LR" have greater discriminating ability, and tests with LRs >10 or <0.1 are very useful in establishing or excluding a diagnosis.

Discussion

The current study developed a machine learning system for radiological diagnosis of digital chest X-ray images. The dataset utilised for training the machine were taken from NIH (USA), whereas the unseen data used for validation were taken from Bangladeshi patients. Initiatives can be taken to generate standard dataset of our own patients, which can assist in developing more precise machine learning system.

The result of validation shown in table I demonstrates a wide range of case-specific variation in accuracy, sensitivity and specificity. The highest accuracy was for diagnosis of emphysema (94.84%), highest sensitivity was for diagnosis of "no findings" (88.25%), which means it would be a good investigation for exclusion any of pathology. However, the remarkably low sensitivity in specific diagnoses (hernia, pleural thickening, edema, pneumonia) can be explained by less number of true positive diagnoses in those special cases. Highest specificity was for diagnosis of pleural thickening (97.23%) and emphysema (96.79%). In future studies, findings of tuberculosis, COVID and some common findings in perspectives of Bangladeshi patients can be brought forward to get the expected results more likely to reality.

In a systematic review performed on 208 articles regarding the use of artificial intelligence for radiological classification of diseases from digital chest X-ray images, it was found that the most widely used datasets were GitHub repository, hospital-oriented datasets, and Kaggle repository. The most considerable value of accuracy, sensitivity, specificity, and area under the ROC curve was reported for ResNet18 in reviewed techniques; all the mentioned indicators for this mentioned network were equal to one (100%). This review revealed that the application of artificial intelligence can accelerate the diagnostic process of COVID-19 [20]. In another meta-analysis assessing diagnostic accuracy for deep learning in medical imaging, the area under curve ranged between 0.864 and 0.937 for diagnosing lung nodules or lung cancer

on chest X-ray or CT scan. The authors suggested a need for the development of artificial intelligence-specific EQUATOR (Enhancing the Quality and Transparency of health Research) guideline regarding key issues in this field[21].

Conclusion

This experimental study developed a machine learning system for radiological diagnosis of digital chest X-ray images with high accuracy, very high specificity and low sensitivity, with case-specific variations in result. Application of this new machine learning system might enable a rapid & accurate radiological diagnosis, rule out any disease in normal chest X-ray, minimise dependency on manpower & logistic, minimise human made error, assist in handling pandemic and unveil the opportunities for future experimental studies. However, this machine learning system is never meant to replace expert human opinion and it can never think beyond the box.

Conflict of interest

The authors have no conflict of interest to declare.

Ackowledgement

This study was funded by Bangladesh Society of Medicine.

References

1. Sikchi SS, Sikchi S, Ali MS. Artificial intelligence in medical diagnosis. *Int J Appl Eng Res*. 2012. doi:10.7326/0003-4819-108-1-80
2. Amato F, López A, Peña-Méndez EM, Vaòhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. *J Appl Biomed*. 2013. doi:10.2478/v10136-012-0031-x
3. Van Ginneken B, Setio AAA, Jacobs C, Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: *Proceedings - International Symposium on Biomedical Imaging*. ; 2015. doi:10.1109/ISBI.2015.7163869
4. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*. 2017. doi:10.1148/radiol.2017162326
5. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017:3-9. doi:1711.05225
6. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8/ : Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. :2097-2106.
7. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Med Phys*. 2016. doi:10.1118/1.4967345
8. Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliveira JL, Lopez MAG. Convolutional neural networks for mammography mass lesion classification. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. ; 2015. doi:10.1109/EMBC.2015.7318482
9. Shin H-C, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. 2016. doi:10.1109/CVPR.2016.274
10. Que Q, Tang Z, Wang R, et al. CardioXNet: Automated Detection for Cardiomegaly Based on Deep Learning. *2018 40th Annu Int Conf IEEE Eng Med Biol Soc*. 2018;2018:612-615. doi:10.1109/EMBC.2018.8512374
11. Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: Detection of findings and presence of change. *PLoS One*. 2018. doi:10.1371/journal.pone.0204155
12. Gacek A, Pedrycz W. *ECG Signal Processing, Classification and Interpretation*.; 2012. doi:10.1017/CBO9781107415324.004
13. Agrafioti F, Hatzinakos D, Anderson AK. ECG pattern analysis for emotion detection. *IEEE Trans Affect Comput*. 2012. doi:10.1109/T-AFFC.2011.28
14. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*. 2017;284(2):574-582. doi:10.1148/radiol.2017162326
15. Shin HC, Roth HR, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging*. 2016. doi:10.1109/TMI.2016.2528162
16. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018. doi:10.1371/journal.pmed.1002686
17. Ayana G, Dese K, Dereje Y, Kebede Y, Barki H, Amdissa D etal. Vision-Transformer-Based Transfer Learning for Mammogram Classification. Diagnostics (Basel). 2023. 13(2):178. doi:10.3390/diagnostics13020178.
18. Ciompi F, Chung K, Van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep*. 2017. doi:10.1038/srep46479
19. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, 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, CVPR, 2017.
20. Rezayi S, Ghazisaeedi M, Kalhori SRN, Saeedi S. Artificial Intelligence Approaches on X-ray-oriented Images Process for Early Detection of COVID-19. J Med Signals Sens. 2022;12(3):233-253. doi: 10.4103/jmss.jmss_111_21.
21. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, Ashrafian H, Darzi A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65. doi: 10.1038/s41746-021-00438-z.