Effective Utilization of Multiple Convolutional Neural Networks for Chest X-Ray Classification

Ravidu Suien Rammuni Silva1 · Pumudu Fernando2

Received: 9 January 2022 / Accepted: 26 August 2022
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2022

Abstract
Out of the numerous types of Medical Imaging modalities available, radiography stands out a bit more than others due to its capabilities of diagnosing diseases and conditions, including life-threatening conditions. Its affordability is another main reason for its prevalence. Chest Radiography holds even higher importance, as it focuses a critical area of the human body. However, interpreting a Chest Radiography image can be challenging and usually done by an experienced Radiologist for accurate results. There are two main issues related to this. One is that in some countries, experienced Radiologists are scarce. The other issue is the inevitability of human errors in diagnoses. Researchers attempt to use Artificial Intelligence to address these two issues. Most of the existing work incorporates Convolutional Neural Networks for this purpose. This paper presents a novel way of parallelizing multiple architectures of Convolutional Neural Networks focusing on Chest X-ray classification. The paper further presents a comprehensive evaluation of the existing architectures with the parallelized results of them using our method. We used four large-scale datasets, including a non-medical one, for the evaluation of our models. We managed to achieve better accuracy for 9 out 13 and 11 out of 14 labels on our two main evaluation datasets. The paper concludes by presenting the limitations and future improvements possible for the system.

Keywords Convolutional neural networks · Deep learning · Medical AI · Medical diagnosis · Radiography

Introduction

Background

Imaging for medical purposes (Medical Imaging) involves various technologies and processes that capture different areas/ organs of the human or the animal body. These imaging modalities contain various types of visual representations corresponding to different structural and qualitative properties of the scanning area. These are very helpful in identifying and confirming a disease’s presence or keeping track of the progression of an already diagnosed disease. Radiography, Magnetic resonance imaging (MRI) and Ultrasonography are typical examples of Medical Imaging Technologies.

A Report by UNICEF in 2019 [1] shows that a child dies because of pneumonia every 39 s. It further shows that pneumonia causes more deaths, especially in children more than any other infectious disease. Tuberculosis is another dangerous disease like that. Even though deadly diseases like Pneumothorax have a comparatively low death rate, [2] shows that it has a high recurrence rate which can be very harmful. Most chest-related diseases can be controlled or sometimes completely cured if identified early and treated well. The most used method in diagnosing chest-related diseases is Chest Radiography. The advantages of chest radiography are the low cost and its convenient operation. Nevertheless, as shown in [3], confusions can occur when diagnosing chest-related diseases using chest radiography due to the complex structural representations in chest radiography images, which can differ from one person to another.

Part of this work has been submitted as a partial requirement for the degree BSc. (Hons.) in Computer Science at the University of Westminster, UK (in collaboration with Informatics Institute of Technology, Sri Lanka).

* Ravidu Suien Rammuni Silva
ravidus.ac@gmail.com; w16267097@my.westminster.ac.uk
Pumudu Fernando
pumudu.f@iit.ac.lk

1 University of Westminster, London, UK
2 Informatics Institute of Technology, Colombo, Sri Lanka

Published online: 27 September 2022
Problem Definition

A report released by Association of American Medical Colleges (AAMC) in June 2020 [4] predicts that by 2032 there will be a shortage of physicians including radiologists of about 122,000. Similarly, as per a report released in 2019 [5] by the Royal College of Radiologists (RCR) shows that only 20% of all the ‘trust and health’ boards in UK has sufficient radiologists required to perform emergency procedures.

Furthermore, poorly resourced areas and countries may not have quality and well-experienced radiology services. In a study in 2015 [6], Rwanda, a country with a population of more than 12 million, only had 11 active radiologists for the whole country. Comparably, Liberia being a country with more than 4 million population, had only two active radiologists till 2015 as per [7]. Moreover, [8] shows that in the United States, the percentage of active radiologists out of the whole medical professionals have been decreasing in the previous years. Reference [9] shows the effect of the insufficiency of the radiologists in the United Kingdom. It also shows the adverse financial effects of insufficient radiologists by noting the overtime paying and the outsourcing costs that the UK government had to bore.

Considering the information provided in the previous sections, it is conclusive that there is an evident insufficiency of professionals with radiographic knowledge in the world. Also, as radiography screening is highly dependent on the experience of the professionals, a guide or a support system would be very helpful even for the experts.

Motivation

With the advancements in hardware and software technologies, researchers tried to use computing to identify diseases in CXRs. The first attempt for a CAD system was found in the 1960’s [10]. Then with the publicly released datasets like [11, 12] and [13], these types of researches have been boosted recently. However, systems built using datasets like those, are currently at the research level. Although they have performed well to some extent, issues like generalization and practicality are still implied. Another main issue is that the current literature primarily focuses on building a model to get higher accuracy levels but not how it will perform in practical medical scenarios. For example, [12]’s algorithm seems promising with the Area Under the Curve (AUC) scores presented in the paper. However, it has been only tested for few pathologies wherein practical scenarios multiple pathologies may need to be identified in a CXR [14].

Contributions of this Work

This paper presents novel Neural Network Architecture, which will accurately detect a selected set of pathologies in a Chest X-ray. This research will also focus on a thorough re-evaluation and critical analysis of the existing works and to present them to the research community in a sensible manner. Below we list the main contributions of this research.

- A critical evaluation of the existing work related to Computer-Aided Detection systems. (We suggest referring to our initial paper [14] for a more broader and critical discussion on the existing works)
- A requirement analysis for the proposing system via a brief survey study.
- Parallelize Block: A novel way to utilize multiple Convolutional Neural Network (CNN) models, especially in a case of classifying multiple classes, which the said CNN architectures are specifically better at.
- ParallelXNet: A CNN-based model trained on multiple datasets freely available to the community.
- A Conceptual Framework for Parallelizing CNN models.
- A systematic re-evaluation of main CNN architectures, comparing them, utilizing all the three main freely available datasets.
- The issue of generalization of the models, proved quantitatively, which existing literature only have hinted at.

Paper Structure

We write this paper as a follow-up paper for our previous publication [14], addressing a few of the critical gaps there. The paper starts with an introduction, providing the necessary background knowledge. Then in the second section, we concisely discuss the existing work that has been done related to our research. In the third section, we present a brief requirement analysis on our research topic. In the fourth section we present our novel CNN architecture with its mathematical explanation. Section six and seven presents the test results, limitations, and the future work.

A Concise Review of Existing Work

The subject ‘Medical Imaging’ is a vast field that involves various different methodologies in capturing the inside body of living beings. Different imaging technologies output a different set of visual representations, which corresponds to a specific set of physical qualities and behaviors of the observing areas of the living body. Hence, this helps to identify and diagnose diseases commonly linked with the physical qualities and behaviors of the living bodies and their organs. These visual-based modalities allowed Computer Vision and Artificial Intelligence (AI) to support the diagnosis process. We suggest referring to our initial paper [14] for a broader and critical discussion on the existing works.
AI for Medical Diagnosis

In the literature, the first attempt to build a Computer-Aided Diagnosis (CAD) system was in 1963 [12]. It was to support the diagnosis of Lung Cancer from Roentgen Images (a.k.a. X-ray Images). However, with the world’s technology back in the day, it was nearly impossible to build an advanced CAD system to equalize or surpass human abilities. However, with the rapid development of the technologies, those systems have seemingly become possible to implement during the recent past. We also note that being ‘possible’ does not necessarily mean ‘easy-to-build’. AI-based CAD systems should be initiated with a good research foundation. Researchers interested in Medical AI have been working hard to build those type of systems. Lung nodule detection [15], eye disease diagnosis [16], cancer detection [17], and localization are some outcomes due to that effort. Research focused on utilizing Deep Learning to tackle COVID-19 like [18, 19] are some notable research carried out recently.

The study in [20] has evidently shown that incorporating CAD systems for the diagnosis of pulmonary nodules raised the diagnosing accuracy of the radiologist by more than 4% average. We believe further quantitative studies should be done in identifying the effect of utilizing AI for Medical Diagnosis and its advantages.

AI for Chest X-Ray (CXR) Diagnosis

As per our initial paper [14] we mainly identify three components that we should consider when discussing an AI-based CXR diagnosis system.

- CXR Labelers
- Image Preprocessing
- AI algorithm; Below we briefly look into each of these component.

CXR Labelers

The result of a CXR diagnosis by a radiologist commonly comes as a narrative report called ‘Radiology Report’. Radiologists briefly write their observations in the report intending for another specialized medical professional or a physician to get a good knowledge of what is being displayed in the CXR image. As per [21], the sections may include the following; Title, Indication, Procedure, Findings, Impression and Footnotes. It further points out a survey that shows one out of two physicians directly looks only into the ‘Impressions’ section of a radiologist’s report implying that section would have higher importance when understanding a radiology report.

Commonly, labelers use Natural Language Processing (NLP) techniques on top of the Text Mining techniques to extract information out of the radiology reports. Labelers [22, 23] were built focusing more on clinical narrative-styled contents and biomedical publications, including Patient Discharge summaries, Radiology reports and Electrocardiograms (ECGs). As noted in [22], it is more biased and performs well for medical texts like Patient Discharge Summaries. However, [22] is a Machine learning (ML) based tool where [23] is built using Ontology-based methodologies. Reference [11] have combines both ML and Ontology-based methods and have quantitively shown an increment in accuracy.

However, none of the mentioned methods were capable in identifying the sense of ‘uncertainty’ which is quite frequent to find in radiology reports as per [24]. They were able to adapt the method in [23] to detect possible ‘uncertainty’ implications in radiology reports. Later, [12] and [13] introduced even advanced and better-performing labelers to label their own CXR datasets.

Image Preprocessing

Similar to the general DL-based classification models, Image Preprocessing techniques also play a major role in this use case. This has become essential for a DL-based Image Classification system as the images we collect to train the models does not always come precisely in the form we want. As the models see image datasets as typical numerical data distributions, it is essential to make sure the data distribution is in a good form which makes easy for the model to get used to and learn from it. Below we list out the main techniques that are related to this research problem. However, this paper does not explore these techniques in-depth as we do not focus much on improving them but on utilizing them.

- Cleaning
- Enhancement
- Segmentation
- Augmentation

Reference [25] proposes a comprehensive way to do data cleaning identifying the image samples that are out of the distribution. In the literature, image segmentation related to CXR images was done mainly in three ways; Rule-based, ML Based and DL Based. References [26] and [27] are examples for some notable work done related to this. Reference [28] presents reliable results segmenting the lung area of CXR images nearing human-level performances.

AI Algorithm

AI Algorithm can be called the core of any AI-based system. Back in the day, researchers used ML-based algorithms in common. However, with the technological improvements and the availability of large-scale datasets, DL-based algorithms easily surpassed ML-based algorithms. Our paper [14] contains a more
descriptive discussion on ML-based algorithms and their limitations. In this paper, we briefly discuss only the notable existing work done using DL-based approaches.

It is significant that the evidence found in literature repeatedly prove the ability of DenseNet architecture [29] and also the ResNet architecture [30] for accurately identifying specifically the pathologies present in Chest X-rays. After an in-depth investigation, we identified that the identity mappings of them are the main contributing factor towards this behavior. These identity mappings evidently work very well in fitting CXR images which have structurally very similar visual model across all the samples. This is due to the visually consistent structure of the human upper skeleton. Even though CNN architectures are capable of learning just by image data and their related labels, evidence like [31] in the literature suggests they perform better when we influence them positively to learn on the CXR images. Below we briefly discuss exiting approaches directly related to our research. Most of these research use datasets in [11] and [12] except for the work done in [32], which uses the largest openly available dataset presently, introduced in [13].

- Reference [33] utilizes the squeeze and excitation blocks (SEnet) published in [34] into their Neural Network. They have used a multi-map transfer layer and max–min pooling methodology to deal with different pathologies separately. However, for few diseases like Hernia, AUC scores have been reduced due to the usage of SE blocks, max–min pooling and multi-map layers. Also, test time augmentation (TTA) techniques have been heavily used, generating six augmentations from each test image sample. Little to no evidence has been depicted on the quantitative effect of these TTA techniques on the final test results evaluation.

- Reference [31] have incorporated lung segmentation methodologies to force the CNN network to focus more on the lung area in the predictions making process. However, little to no improvement can be seen in AUC scores, even with heavy computations compared to [35] that uses the DenseNet-121 network with very few modifications.

- Reference [36] proposes a branched architecture with three main branches. Firstly, a CNN network tries to localize the abnormality. Then that area is passed through again in a CNN network. Then the detection results taken from both the branches were combined to get the final result. However, this makes the whole architecture highly dependent on the CNN network that tries to localize the abnormality. Most importantly, even though the paper presents a state-of-the-art result, they have used a custom split for the evaluation. This makes the results incomparable to the other results found in the literature. Also, the split does not guarantee that the patient overlapping would not be present between the splits.

- The approach presented in [37] is currently at the 2nd position in the CheXpert [38] competition. They have used model ensembling and Label Smoothing Regularization (LSR) in their approach. They have also used a novel approach called Conditional Training (CT) methodology, which uses hierarchical dependencies among different pathologies to each other to improve the predictability of the model. However, the evaluation of the model was only done on five pathologies out of twelve found in the training split of the dataset. Furthermore, the authors have acknowledged the possible inabilities of the CT method to generalize well on different distributions. Hence, this would possibly make the detection results prejudice to the distribution that training data is originated from.

- Work [32] have designed a CNN-based architecture specialized for CXRs to take the two views of one of them, ‘frontal and lateral’, and work with them concurrently for making the final detection result. However, the presented AUC scores for each pathology are relatively low.

Most of the CXR datasets used by the aforementioned research studies were compiled from geographically neighbor sources. Due to this fact, the models trained on those datasets may not perform up to the exhibited evaluation results in reality. This could be due to hidden and unwanted patterns due to the equipment used, medical conditions of the belonging patients, followed imaging procedures and technologies. Furthermore, almost all the existing works merely rely on the CXR image (or more views whenever possible) to detect pathologies on CXRs. However, in reality, medical professionals consider various biological factors, including but not limited to; symptoms, age, gender, geographical environment, behavioral patterns, medical history and various other clinical and non-clinical tests.

**End-to-End Systems**

Table 1 shows similar systems currently available. It can be seen, even though there are clinically certified systems for scanning reports like ultrasound scanning and CT scanning [39], CADs for CXR are still under very preliminary stages and have never been ‘certified clinically’ for medical usage. Although attempts like [40] show good accuracies in the publications, the datasets used are not made available publicly and it is unclear whether the algorithms are actually effective or whether that the ‘amount of data’ used to train or test is what causing the good accuracy scores.

**Survey Study**

**Questionnaire**

A Questionnaire was distributed among a limited, targeted audience. The complete questionnaire is available in the appendices. The targeted audience was the Medical Students in their final study year (majority) of MBBS degree in Sri Lanka from different state universities. Fifty-three participants have participated in this preliminary survey. Below is the analysis of the results.
of the questionnaire. As interviews ("Interviews") were focused on the field experts, the questionnaire was circulated among the practicing medical students to get a better understanding of high-level requirements and the comments on them for the proposing system. This also helped us to understand how it will support medical education and in a practicing environment. See Appendix for the complete Questionnaire results.

The majority (77%) of the responses come from final year (fifth year) students following an MBBS degree. They are the most targeted audience in the questionnaire as generally, they are the most experienced medical students.

- It could be seen that, very often, medical students get to diagnose patients by their CXR. However, with the responses to the 5th question, it is clear that for further confirmation, a more specialized medical test might be necessary in most cases to confirm their findings. This was also noted in the interviews.

- Most of the respondents disclosed that they further refer to a specialized doctor or a colleague to get their findings to confirm. Hence the system might potentially help as a support system.

- From questions 7 and 8, it can be stated that even though they are optimistic about the system, at the same time, they are uncertain whether the system would be adequate to overcome their errors. Hence the system must be designed in a way to gain their confidence having a provable accuracy.

- From the 10th and 11th questions, a clear majority expresses the importance of accuracy in CXR screening over its urgency.

- From the general comments of the questionnaire, the majority seems to have a positive thought about the system. However, one respondent (medical student) has suggested that the system might be helpful for radiologists but can be misleading for physicians. Further, the respondent suggests that it is better not to give medical students the system as they should learn to screen CXRs by their own. Some other few comments were made by medical students conveying the need for more pathologies to be added to the detection list of the system.

### Interviews

We also conducted interviews as a part of the requirements engineering process of this research. Interviews were explicitly done on a targeted audience to get medical-related knowledge, understanding and the requirements at a concrete level. These were dedicated to the doctors to investigate their requirements on a more profound level. The questionnaire was dedicated to the medical students to capture requirements at a higher level but from a larger audience.

Interviews were conducted in a semi-structured way complying with the questionnaire questions. The selection of the interviewees was made considering their area of expertise. Two doctors with high experience, specifically in thoracic disease diagnosis and one doctor with high experience in the research field were chosen for the interviews. We conducted these interviews via online video conferencing and we spent approximately 20 min on each interview.

Analysis of the interviews done are as below,

- Chest radiography is very commonly used, especially in hospitals dedicated to thoracic or other respiratory system-related diseases. Both the doctors mentioned that they take a Chest X-ray (CXR) from almost every patient admitted to the hospital as a standard procedure.
- All the interviewees believe that having an automated detection report of the potential findings in a CXR would help proceed with or to confirm the interpreted results from CXR screening further.
- All interviewees believe that the system would be even more helpful if it could detect more types of diseases on CXR as numerous types of abnormalities could be present in a CXR.
- All interviewees prefer to use the system through a mobile device.
- Also, interviewees think that these systems may not be able to replace human radiologists completely.
- One interviewee suggested that the accuracy rates must be analytically evaluated, and the error margins must be pre-
cisely exhibited in the system interface and in the final detection report to avoid over-dependence on the auto-generated detection report.

ParallelXNet

Why Parallelizing?

Almost all the existing literature uses some sort of an ensembling method in their method. When thoroughly investigating the existing literature related to pathology detection in CXRs, various architectures have been seen to be used for the purpose. An extensive discussion has been driven on these in ("A Concise Review of Existing Work"). Also, our own publication [14] also has some critical discussions. Those discussions show that different neural network architecture performs differently in detecting each different pathology as numerous visually different pathologies could be found using CXRs. Hence our goal was to make the best out of all the architectures.

For the above purpose, a commonly used technique is the Ensembling Method, which basically implies the averaged output of multiple neural networks that have been trained separately. But this method does not guarantee “the best from each” neural network’s that is ensembled together as some outputs might get neutralized in terms of accuracy while the averaging process. For this purpose, we suggest a novel block that can be used, instead of averaging the outputs after the training process.

Parallelize Block

The illustration in Fig. 1 shows the inner connections of the block we propose. This design of the block is mostly motivated by the design behind SE Blocks [34]. As the figure shows, the block has its own mini-neural network embedded into it that learns which channel(s) out of all the neural networks is more important at predicting a given disease. Then this numerical representation of ‘importance’ is used to weight each channel, amplifying the channels that are important. This would highlight the important features out of all the variants of same feature set, identified by different neural networks, which ultimately helps achieving the “best from each” concept.

Theory

The Channel Attention based Parallelizing (CAP) block tweaks the contribution of each channel of the final convolutional layer from each of the parallelizing neural network to the final output of the overall model.

Let the inputs for the block be \( \{X_1, X_2, X_3\} \in \mathbb{R}^{H \times W \times C} \); where \( C \) can be varied. \( W \) and \( H \) are matched padding with zeroes. Also let \( C_1, C_2, C_3 \) are channel counts of the inputs respectively. Then, concatenating the input we get \( X \in \mathbb{R}^{W \times H \times C} \);

\[
C = C_1 + C_2 + C_3. \quad (1)
\]

Also, \( X = \{x_1, x_2, x_3, \ldots, x_C\} \) where \( x_n \) represent the \( n \)th channel. Then the concatenated input is average pooled to get \( y_{ca} \in \mathbb{R}^{C \times 1 \times 1} \) as below (2)

\[
y_{ca} = \frac{1}{W \cdot H} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \quad (2)
\]
\( y_{ca} \) here will be the input for the embedded fully connected neural network (FCN). FCN is involved in learning the contribution from each channel out of all the channels from all the parallelized neural networks. The FCN behaves as per the below equations.

Let the learning functions of layer FC1 and FC2 be \( \mathcal{F}_{FC1}, \mathcal{F}_{FC2} \) respectively. Then the final output of the FCN, \( Z \);

\[
Z = \sigma \left[ \mathcal{F}_{FC1} \left( \delta \left( \mathcal{F}_{FC1} \left( y_{ca} \right) \right) \right) \right] \\
Z = \sigma \left[ w_2 \cdot \left( \delta \left( w_1 \cdot y_{ca} + \beta_{FC1} \right) \right) + \beta_{FC2} \right]
\]

where \( w, \beta, \delta \) and \( \sigma \) are learned parameters, bias terms, ReLU activation function and the sigmoid function respectively. Note that the dimensionality is reduced by a factor of \( \tau \), which results,

\[
w_1 \in \mathbb{R}^{C \times \tau}, w_2 \in \mathbb{R}^{C \times \tau} \text{ and } \\
\beta_{FC1} \in \mathbb{R}^{\tau \times 1}, \beta_{FC2} \in \mathbb{R}^{\tau \times 1}.
\]

Above \( Z \) represent a vector that learns the significance of each channel that is concatenated in \( X \). When properly learned, each element would represent a numerical and quantitative representation for each channel’s ‘importance’ towards the accurate prediction of the final result. To make use of this equation, the \( Z \) vector is then used to weight the channels in \( X \). Then the new weighted input \( X_w \) is given by (5)

\[
x'_w = z_n \cdot x_n ; \quad (\forall n \in \mathbb{Z}^+ \land n \leq C)
\]

where \( x_n = \{x_1', x_2', x_3', …, x_C'\} \) and \( Z = \{z_1, z_2, z_3, …, z_C\} \).

The new weighted input \( X_w \) is again average pooled similarly to Eq. 6 and then connected to the final output dense layer as shown in Fig. 1.

**ParallelXNet: Architecture Design**

The diagram in Fig. 2 shows the architecture of the ParallelXNet neural network. It contains a common input layer which then connects to the three CNNs. The choice of these three specific architectures was made by investigating their performance, especially in the field of pathology detection in CXRs.

We strongly suggest that this can also be used for a variety of applications other than pathology classification in CXRs, which are out of scope for this research. Also, this suggests that various combinations of CNN architectures can be used within this architectural framework proposed depending on the requirement.

**Data Preparation**

Mainly [11] and [13] datasets were used for this research. Before training, all the datasets were split according to the provided official split. This allowed a fair comparison with all the models and made sure patient overlapping is not present among the splits. As mentioned in previous sections, multiple datasets were used for the development of this system. All the images were resized to 320×320 px resolution. The Batch size was set to 16. To avoid unwanted biases and make the image data easy to learn from, sample wise normalization was used in all splits. Data was augmented in train split by randomly flipping the CXR images while training. The flipping was only done along the horizontal axis. This artificially provides more data for the neural network to train on, ultimately improving the model accuracy.

As no dataset is perfectly balanced, the datasets that we acquired also had this issue of imbalance. Due to this, the loss function will be more affected by negative labels rather than positive labels. Hence, it will make the neural network more biased for the learning of the label type with a higher count when training. To solve this, we used weighted Binary Cross Entropy Loss as the loss function.

\[
-\frac{1}{N} \sum_{i=0}^{N} \left( \beta Y_i \log \hat{Y}_i + (1 - \beta) (1 - Y_i) \log (1 - \hat{Y}_i) \right)
\]

where \( \beta \) is the weighing constant, \( Y \) is the ground truth label and \( \hat{Y} \) is the predicted label. \( \beta \) is the vector containing the frequencies of negative samples per each class. \( N \) represents the batch size here as we calculate loss per batch.

**Training**

While training, weights were saved after each epoch. We reduced the learning rate by a factor of 0.1 after training another epoch with the same learning rate when an increase of the validation loss was detected. The minimum learning rate was capped at 0.000001.

Chosen hyperparameters were as per below; these were selected after numerous pre-tests and considering the evidence from the existing literature.

- Step per epoch: 4500 [Batch Size: 16]
• Learning Rate: 0.0001
• Epochs: MIMIC-CXR Dataset-18 | ChestX-ray14 Dataset-9

In summary, for the MIMIC-CXR dataset, we trained the neural network more than 3.5 times, and for the ChestX-Ray14 dataset, 14 times over the whole dataset. Training using the MIMIC-CXR dataset was done in two iterations due to the notebook’s runtime time limit limitations. One Nvidia Tesla P100 16 GB GPU was used in this research for the model fitting purposes.

Web Application

Development Languages

We selected Python [v3.7] as the primary language for the development of the model and the Server of the Web Application. It is an increasingly popular language, especially for Machine Learning and Deep Learning based development work. Python is a language with a very clean and natural syntax yet a very powerful language that allows developers to develop their application following either an object-oriented or a structural approach. The choice of Python is also motivated by its support for the Jupyter Notebooks, which allows the model development to be done more visual and explanatory way. Also, the language supports a variety of frameworks and libraries. These features make the language a perfect entry point for a beginner but who wants to produce something innovative.

Development Frameworks

We used a Python-based framework, Flask [v1.1.2], to develop the back end of the application. The main reason for this choice is the library support it has to the deployment of the deep learning models. The framework has a very easy-to-use simple structure. We firmly believe that Flask is one of the best frameworks to deploy a deep learning model. Flask also provides a very convenient way of developing RESTful APIs based on Python.

Development Architecture

3-Tiered Architecture was selected as a guide in building the system. The functionalities that fulfill the requirements of the system are meticulously placed among the three tiers. Below we discuss briefly about each main tier.

Data Tier

This tier represents the data that is needed for the system to function. The Deep Learning Model, which is the most important component in the system, is placed in this tier. Also, the localized and the uploaded images live in this tier. This saves computational power by eliminating any redundant localizations. The symptoms related to each disease are stored at the server’s side, allowing future changes and improvements.

Logic Tier

This tier contains all the necessary tools need to perform the computations and finally come up with the required output. Sub parts of this tier are as follows.

- Detection Engine: Carries out the detection of pathologies on the inputted Chest X-Ray and produces with a probability percentage.
- Localization Engine: Localizes the area the potential area that the detected pathologies might exist on the CXR. This is done by generating a heatmap and superimposing it on the original CXR image.
- Presentation Engine: Carries out the necessary functionalities needed to present the data generated in the detection engine in an informative way onto the presentation tier.

Presentation Tier

The data generated in the logic tier is processed into a meaningful representation in this tier. These processed results will be then sent to the front end to be displayed to the user when requested. This tier is the interface for the front-end application to the server functionalities.

TensorFlow and Keras

We used TensorFlow-based version of Keras [TensorFlow v2.4.1] to develop the deep learning model. Keras is an easy-to-use, high-level framework that allows developers to build, train and test ML models with a minimum number of code lines. Alternatives like PyTorch was also considered. However, considering the objectives and the requirements of this project, we selected the Keras with TensorFlow backend. Although Keras is a high-level library, its integration with TensorFlow allows the developers to do development at the low level of the library while making the best use of the high-level functionalities. This makes the combination of TensorFlow and Keras beneficial, mostly for a beginner who wants to produce novel and innovative model algorithms.

Evaluation

We consider the evaluation of this system very important, especially as the end system is intended to be used in a medical based environment dealing with humans. Due to this criticality of the system, we first defined our objectives behind the evaluation, which then was used to construct the test goals. The principal testing objectives are as follows.
• Certify that the system performs its processes and produces the outputs as expected.
• Certify that the ‘algorithm’ and the ‘system’ meet the expected Accuracy, Quality and Standards when performing its processes.
• Perceive the reaction and behavior of the system in unexpected situations.

As mentioned earlier, these defined objectives were utilized to define the below test goals.

• Goal 1: Investigate the quantitative and qualitative effects of the proposed algorithm.
• Goal 2: Investigate the rate of achievement of the defined functional requirements of the system.
• Goal 3: Investigate the rate of achievement of the defined non-functional requirements of the system.
• Goal 4: Investigate any unexpected behaviors of the system.

In this paper, we focus more on Goal 1 with the aim of portraying the capabilities of the proposed algorithm.

Test Setting

Giving attention to the facts discussed about the accuracy metrics to be unbiased and objective as much as possible, we fixed the below settings for the algorithm.

Dataset Split

Parameter Setting

Testing of the model was done extensively under two datasets, as in Table 2. We used one Nvidia Tesla P100 16 GB GPU for our project (see Table 3 for full specification). Only the official splits were used for the tests. That made sure no patient overlapping exists between the splits. For the MIMIC-CXR dataset, we used the NegBio label set due to its better accuracy. For the final model, ‘ParallelXNet-64’ (P-64) and ‘ParallelXNet-128’ (P-128) was used setting \( \tau = 64 \) and \( \tau = 128 \) respectively. These were selected after a series of repetitive train-validate cycles with the validation split. We exposed the test split to the model only at the final test inference run, after freezing the hyperparameters. We also suggest that further fine-tuning can be done on these parameters.

For the MIMIC-CXR, we took weights from the final three epochs after 18 epochs when testing. For the ChestX-Ray14, we trained nine epochs, and the final three’s weights were taken. The six models were then ensembled to get ‘P-Ens’. Test Time Augmentation (TTA) was also used flipping each test sample horizontally. Apart from those mentioned settings, all the other settings were maintained equivalent for the test.

Table 2 Dataset split specification

| Dataset     | Train split | Validation split | Test split |
|-------------|-------------|------------------|------------|
| MIMIC-CXR   | 368,945     | 2991             | 5159       |
| ChestX-ray14| 86,524      | 25,596           | [Val. split] |

Table 3 Hardware Setting

| CPU         | Intel xeon |
|-------------|------------|
| OS          | Linux-5.4.89 + -x86_64-with-debian-buster-sid |
| GPU         | Tesla P100-PCIE-16 GB |
| RAM         | 16 GB      |
| Hard disk   | 20 GB      |

AUC Scores

As shown the Tables 4, 5, Test Time Augmentation has been able to increase the accuracy of the model. Although the increment is not significant, it is consistent throughout the datasets.

Benchmarking

We trained the existing model architectures providing the exact same conditions as in previous tests. Dataset splits and the labels are also the same. The only change done was the model architecture. These conditions and results can be verified by contacting the us, through the model training Jupyter Notebooks we developed. Also see our GitHub repository [https://github.com/SuienS/parallexnet-cxr-classifier]. Used architectures are,

• ResNet50v2 (R-50v2) [30]
• DenseNet-121 (D-121) [29]
• DenseNet-169 (D-169) [29]
• Ensemble of above three, each with the ensemble of the final three epoch weights (R-D-Ens)

The design of the ‘Parallelize Block’ is discussed in (“ParallelXNet”). We introduce this as a better alternative for conventional Ensembling of the Models. This is a solid verification (see Tables 6, 7) because of the consistent results on both the datasets used. Mean values respectively: 0.7885, 0.8246. We strongly suggest this method to be implemented using various architectures available, considering their abilities to detect different pathologies. Also, the number of parallelized architectures can also be increased subjected to the hardware capabilities.
Table 5  AUC scores when tested on ChestX-ray14 dataset

| Model pathology      | P-64   | P-128   | P-Ens  | P-Ens (with TTA) |
|----------------------|--------|---------|--------|------------------|
| Nodule               | 0.7826 | 0.7807  | 0.7875 | 0.7894           |
| Cardiomegaly         | 0.8901 | 0.8927  | 0.8958 | 0.8967           |
| Emphysema            | 0.9294 | 0.9312  | 0.9335 | 0.9348           |
| Fibrosis             | 0.8321 | 0.8344  | 0.8381 | 0.8395           |
| Edema                | 0.8502 | 0.8474  | 0.8526 | 0.8534           |
| Consolidation        | 0.7529 | 0.7527  | 0.7576 | 0.7594           |
| Pneumonia            | 0.7386 | 0.7353  | 0.7411 | 0.7413           |
| Atelectasis          | 0.7863 | 0.7823  | 0.7888 | 0.7892           |
| Pneumothorax         | 0.8720 | 0.8740  | 0.8773 | 0.8786           |
| Effusion             | 0.8359 | 0.8370  | 0.8399 | 0.8406           |
| Mass                 | 0.8329 | 0.8414  | 0.8433 | 0.8435           |
| Infiltration         | 0.6984 | 0.7028  | 0.7041 | 0.7051           |
| Hernia               | 0.8742 | 0.8905  | 0.8911 | 0.8928           |
| Pleural thickening   | 0.7897 | 0.7889  | 0.7942 | 0.7952           |

Training Performance Testing

Note that for training times in Table 8, hardware is as per Table 3 and training settings and hyperparameter choices are the exact same for all cases. Also, note that the file type that the model is saved is .h5. It saves the whole model in a single file. But further compression mechanisms could be incorporated to reduce model size.

Outcomes from the test show that ResNet50v2 has the shortest training time. Outcomes further show that the DenseNet-121 model has the smallest model size. Although evidently, the ParallelXNet is architecture proven to be better at accuracy, for mobile-based hardware, we recommend using DenseNet-121 due to its significantly low model size. We suggest that ResNet50v2 could be better for re-trainable systems as it has the lowest training time. We also acknowledge that the training time could mainly vary depending on the below key aspects,

- Hardware Performance (especially GPU and RAM), ML Library used and Training images' sizes and the file type.

Generalization Test

The test setting for the generalization test was set to the same as previous algorithm tests. Additionally, we trained ParallelXNet-128 on both MIMIC-CXR and CheXpert datasets. Out of 18 epochs, 9 were trained under MIMIC-CXR, and the latter 9 epochs were trained on CheXpert [12] by Stanford University, US. It contains 224,316 labeled CXRs associated with 65,240 imaging studies. The purpose of this test is to identify how the trained model generalizes between datasets. We note that generalization is mostly affected by the trained set of data and not by the model’s architecture. We also note that although there are 5159 images in the test set of the MIMIC-CXR dataset, only 234 images were given as the test set in CheXpert. This may also have some effect on the test results.

The results in Table 9 show that all the labels other than for one, have good stability with more generalized training. The drastic difference in the ‘Enlarged Cardiom.’ Label is suspected of having caused by the ill performed labeling of the CheXpert labeler, which is also quantitatively shown in [13]. We also emphasize that although some pathologies show an increased AUC, some pathologies have not. This could be due to the difference in the labeling performances, but this also suggests that more data does not always mean more accuracy. Another finding that we could derive from these test results is that the AUC score is not an absolute representation of the accuracy of the model. It represents how well the model can perform on a specific ‘trained distribution’.

We also note that the results are not consistent when captured images from a camera due to various light conditions and camera quality. As also noted in the existing literature, these can be caused by various factors of the camera that is used to capture the CXR. Table 10 shows an example of that. It shows that the images’ technical specifications could significantly affect the inference results of the model even though the images look visually identical to the human eye vision.

Additional Evaluations on the Model

Furthermore, we tested our model also on the CIFAR-10 dataset to further verify the applicability and performance of it in non-medical scenarios. For the results in Table 11, we only used the official splits for the training and evaluation of the models. Also, note that we used SoftMax activation for the final Dense layer as the data samples are strictly single class. Mathews Correlation

SN Computer Science
Coefficient (MCC) introduced in [43] was also used here to solidify the accuracy results. These results suppose that the parallelization technique that we introduced in this paper can be used instead of typical model ensembling for non-medical use cases as well. ‘R-D-Ens’ and ‘P-Ens’ were trained similarly to it of in (“Training Performance Testing”).
A Critical View on Our Test Results

Parallelization of the CNN Architectures through the introduced ‘Parallelize Block’ performs consistently well across the two datasets used. This proves the effectiveness of the introduced block with solid evidence. The additional test we conducted on the CIFAR-10 dataset further solidifies it. Also, we note that the improvement in the AUC score is limited to 1–2% per disease. However, as the model is intended for medical usage, any improvement could be considered a valuable improvement. Also, our parallelized architecture was evidently seen resilient for different data distributions than others.

Further, we strongly suggest that the introduced parallelized architecture framework can be used with different combinations of CNN architectures which will potentially increase the accuracy scores furthermore.

Limitations and Future Work

- Below we acknowledge the predominant limitations of this research.
- The models built according to the algorithm introduced have larger file sizes and require high-end hardware to deploy. This might cause scalability issues for a production-level environment.
- The built model is only trained on the 14 pathologies which are available in the datasets.
- Due to the time restrictions of the Kaggle Notebooks, continuous training was not possible, which led to training the model in parts.

We find below features as possible features that could be implemented and integrated into the built system without the need of any significant modifications.

- In-system model re-training: This will allow overcoming the issue of generalization. ‘Federated Learning’ based approach could be implemented to enable convenient and effective retraining.
- Custom model uploads: This allows users to train models and upload, making the web app universal.
- A functionality to allow medical professionals to investigate the patient’s medical history directly through the system. This will help them in making diagnosis decisions. Also, this can

Table 9 Generalization test results: ‘Lung Lesion’ results of CheXpert test set was neglected due to insufficient test samples

| Model pathology          | Train set                      | Test set                     |
|--------------------------|--------------------------------|------------------------------|
|                          | MIMIC-2020 + CheXpert          | MIMIC-2020                   |
| Enlarged cardiom         | 0.5361                         | 0.5720                       |
| Cardiomegaly             | 0.8149                         | 0.7673                       |
| Lung lesion              | 0.3176                         | 0.7201                       |
| Lung opacity             | 0.9360                         | 0.6738                       |
| Edema                    | 0.9173                         | 0.8239                       |
| Consolidation            | 0.9299                         | 0.7426                       |
| Pneumonia                | 0.6875                         | 0.6864                       |
| Atelectasis              | 0.7948                         | 0.7279                       |
| Pneumothorax             | 0.8695                         | 0.8881                       |
| Pleural effusion         | 0.9337                         | 0.8902                       |
| Pleural other            | 0.8927                         | 0.7816                       |
| Fracture                 | N/A                            | 0.7190                       |
| Support devices          | 0.9650                         | 0.8630                       |
|                          | P-Ens                          | P-Ens                        |
|                          | 0.7107                         | 0.7932                       |
|                          | 0.7192                         | 0.7031                       |
|                          | 0.7597                         | 0.8419                       |
|                          | 0.7372                         | 0.7703                       |
|                          | 0.7579                         | 0.7703                       |
|                          | 0.8706                         | 0.8985                       |
|                          | 0.8466                         | 0.8985                       |
|                          | 0.6916                         | 0.8985                       |
|                          | 0.9085                         | 0.8985                       |

Table 10 Localization result on Digital CXR Vs. Camera Captured CXR. The following sample was taken from MIMIC-CXR 2020 [13]

| Label: cardiomegaly      | Digital CXR | Camera (smartphone) captured CXR |
|--------------------------|-------------|----------------------------------|
| Detection rate           | 83.20%      | 42.44%                           |

Table 11 Overall metrics when tested comparatively CIFAR-10

| Model metric            | R-D-Ens  | P-Ens  |
|-------------------------|----------|--------|
| Sensitivity             | 87.94%   | 88.55% |
| Specificity             | 98.66%   | 98.75% |
| Precision               | 88.00%   | 88.57% |
| Accuracy                | 97.58%   | 97.72% |
| Balanced accuracy       | 93.30%   | 93.64% |
| F1-score                | 87.94%   | 88.56% |
| MCC                     | 0.8663   | 0.8729 |

‘ParallelXNet’ is better in terms of all the metrics considered for CIFAR-10 dataset
possibly allow even the model to consider the past CXRs of a patient when his/her new CXR are being detected through the system.

In conclusion, we should note that our work shows a pathway to develop further the existing DL-based models focusing on medical-related applications. Although the models nor the whole system is not recommended for medical usage, our work can be further developed and trained using larger and more varied datasets to increase the accuracy and reliability. We believe systems like these should always be supported and monitored by a trained professional and only be used as a supportive tool. However, we believe more research work like these would one day replace human professionals.

Appendix
Acknowledgements  We thank N F (MBBS, DFM), HB (MBBS, MD) and PDeS (MBBS, MSc), for their invaluable comments and knowledge contributions throughout this research. A special thanks should go to the Kaggle Notebooks, without which this research would not be possible. A huge appreciation should also go to the creators of the datasets, who has done immense work and released it freely on request to the research community.

Funding  This is a self-funded work by the authors.

Data availability  All the material related to the developed system is available via below link; https://github.com/SuienS/parallexnet-cxr-classifier.

Code availability  The code for the developed system is available via below link; https://github.com/SuienS/parallexnet-cxr-classifier.

Declarations

Conflict of interest  Authors declare no conflicts of Interests.

Ethical approval  The study was carried out under the strict supervision of Informatics Institute of Technology, Sri Lanka and University of Westmin-ster, UK. Since, no direct human involvement was observed in the study, no special Ethical approval was required.

Consent to participate  Informed consent was obtained from all individual participants included in the study including Medical Doctors and Medical Students who participated in the Survey Study relating to this research.

Consent for publication  Informed consent for publishing the results was obtained from all individual participants included in the study including Medical Doctors and Medical Students who participated in the Survey Study relating to this research.

References

1. UNICEF, ‘Pneumonia in Children’, UNICEF DATA, 2019. https://data.unicef.org/topic/child-health/pneumonia/ Accessed 9 Jan 2020
2. Wakai A. ‘Spontaneous pneumothorax’. 2008; 11
3. Porpodis K et al. ‘Pneumothorax and asthma’. J Thorac Dis. 2014; 6, 1, [Online]. Available: http://jtd.amegrroups.com/article/view/2173
4. Association of American Medical Colleges, ‘The Complexities of Physician Supply and Demand: Projections From 2018 to 2033’. Jun. 2020. [Online]. Available: https://www.aamc.org/media/45976/download/attachment

5. Royal College of Radiologists, ‘Clinical Radiology UK. Workforce Census Report 2018’. 2019. [Online]. Available: https://www.rcr.ac.uk/system/files/publication/field_publication_files/clinical-radiology-uk-workforce-census-report-2018.pdf

6. Harvard Medical School et al. Imaging in the land of 1000 hills: rwanda radiology country report. J Glob Radiol. 2015; 1, 1, https://doi.org/10.7191/jgr.2015.1004

7. UMass Memorial Healthcare et al. Diagnostic radiology in libera: a country report. J Glob Radiol. 2015; 1, 2, https://doi.org/10.7191/jgr.2015.1020

8. Rosenkrantz AB, Hughes DR, Duszak R. The US radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology. 2016;279(1):175–84. https://doi.org/10.1148/radiol.2015150921

9. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ. 2017. https://doi.org/10.1136/bmj.j4683.

10. Lodwick GS, Keats TE, Dorst JP. The coding of roentgen images for computer analysis as applied to lung cancer. Radiology. 1963;81(2):185–200. https://doi.org/10.1148/1.821.185.

11. 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, 2017.

12. Irvin J et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, ArXiv190107031 Cs Eess, 2019, Accessed: 24 Jul 2020. [Online]. Available: http://arxiv.org/abs/1901.07031

13. Johnson AEW et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, ArXiv190107042 Cs Eess, 2019, Accessed: 24 Jul 2020. [Online]. Available: http://arxiv.org/abs/1901.07042

14. Silva RRS, Fernando P. An extensive survey of machine learning based approaches on automated pathology detection in chest X-rays’, 2021, 9

15. Setio AAA, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal. 2017;42:1–13. https://doi.org/10.1016/j.media.2017.06.015.

16. Gulshan V, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402. https://doi.org/10.1001/jama.2016.17216.

17. Chen C, Kavuri VC, Wang X, Li R, Liu H, Huang J. Multi-frequency diffuse optical tomography for cancer detection, in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA, 2015, pp. 67–70. https://doi.org/10.1109/ISBI.2015.7163818.

18. Wynants L et al (2020) Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. https://doi.org/10.1136/bmj.m1328.

19. Jamshidi M, et al. Artificial Intelligence and COVID-19: deep learning approaches for diagnosis and treatment. IEEE Access. 2020;8:109561–95. https://doi.org/10.1109/ACCESS.2020.3001973.

20. Kobayashi T. Effect of a computer-aided diagnosis scheme on radiologists’ performance in detection of lung nodules on radiographs, 1996.

21. Wilcox JR. The written radiology report, 2006, 4

22. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform. 2015;57:28–37. https://doi.org/10.1016/j.jbi.2015.07.010.

23. Aronson AR, Lang F-M. An overview of metapain: historical perspectice and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36. https://doi.org/10.1136/jama.2009.002733.

24. Peng Y. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. 2017, 9

25. Çallı E, Murphy K, Sogancioglu E, van Ginneken B. FRODO: free rejection of out-of-distribution samples: application to chest x-ray analysis’, ArXiv190701253 Cs Eess Stat, 2019, Accessed: 08 Aug 2020. [Online]. Available: http://arxiv.org/abs/1907.01253

26. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Comput Vis Image Underst. 1995;61(1):38–59. https://doi.org/10.1006/cviu.1995.0104.

27. Wen H, Qi W, Shuang L. Medical X-ray image enhancement based on wavelet domain homomorphic filtering and CLAHE. 2016, 6

28. Dai W et al. SCAN: structure correcting adversarial network for organ segmentation in chest X-rays, ArXiv170308770 Cs, 2017, Accessed: 02 Aug 2020. [Online]. Available: http://arxiv.org/abs/1703.08770.

29. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks, ArXiv160806993 Cs, 2018, Accessed: 06 Aug 2020. [Online]. Available: http://arxiv.org/abs/1608.06993.

30. He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks, ArXiv160305027 Cs, 2016, Accessed: 29 Nov 2020. [Online]. Available: http://arxiv.org/abs/1603.05027.

31. Liu H, Wang L, Nan Y, Jin F, Wang Q, Pu J. SDFN: segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. Comput Med Imaging Graph. 2019;75:66–73. https://doi.org/10.1016/j.compmedimag.2019.05.005.

32. Rubin J, Sanghavi D, Zhao C, Lee K, Qadir A, Xu-Wilson M. Large scale automated reading of frontal and lateral chest X-Rays using dual convolutional neural networks. ArXiv180407839 Cs Stat, 2018, Accessed: 03 Aug 2020. [Online]. Available: http://arxiv.org/abs/1804.07839.

33. Yan C, Yao J, Li R, Xu Z, Huang J. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. Proc 2018 ACM Int Conf Bioinforma Comput Biol Health Inform. 2018. https://doi.org/10.1145/3233547.3233573.

34. Hu J, Shen L, Albane S, Sun G, Wu E. Squeeze-and-Excitation Networks, ArXiv170901507 Cs, 2019, Accessed: 08 Aug 2020. [Online]. Available: http://arxiv.org/abs/1709.01507.

35. Rajpurkar P et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning, ArXiv171105225 Cs Stat, 2017, Accessed: 05 Aug 2020. [Online]. Available: http://arxiv.org/abs/1711.05225.

36. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification, ArXiv180109927 Cs, 2018, Accessed: 24 Jul 2020. [Online]. Available: http://arxiv.org/abs/1801.09927.

37. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels, ArXiv191106475 Cs Eess, 2020, Accessed: 03 Aug 2020. [Online]. Available: http://arxiv.org/abs/1911.06475.

38. CheXpert: a large chest x-ray dataset and competition, CheXpert: A Large Chest X-Ray Dataset And Competition. https://stanfordmlgroup.github.io/competitions/chexpert/

39. ‘LungLifeAI’, LungLifeAI, 2019. https://www.lunglifeai.com/

40. ‘Qure.ai’, Qure.ai—Artificial Intelligence for Radiology, 2020. https://qure.ai/

41. ‘Oxipit’, Oxipit—The first AI chest X-Ray radiology suite for healthy patients, 2020. http://oxipit.ai/

42. Putha P et al (2020) Can artificial intelligence reliably report chest X-rays?: Radiologist validation of an algorithm trained on 2.3 million X-rays, ArXiv180707455 Cs, 2019, Accessed: 18 Aug 2020. [Online]. Available: http://arxiv.org/abs/1807.07455.
43. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta BBA-Protein Struct. 1975;405(2):442–51. https://doi.org/10.1016/0005-2795(75)90109-9.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.