Observer performance evaluation of the feasibility of a deep learning model to detect cardiomegaly on chest radiographs

Acta Radiologica Open
11(7) 1–11
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20584601221107345
journals.sagepub.com/home/arr
SAGE

Pranav Ajmera[1], Amit Kharat[1], Tanveer Gupte[2], Richa Pant[2], Viraj Kulkarni[2], Vinay Duddalwar[3] and Purnachandra Lamghare[1]

Abstract

Background: Cardiothoracic ratio (CTR) is the ratio of the diameter of the heart to the diameter of the thorax. An abnormal CTR (>0.55) is often an indicator of an underlying pathological condition. The accurate prediction of an abnormal CTR chest X-rays (CXRs) aids in the early diagnosis of clinical conditions.

Purpose: We propose a deep learning (DL)-based model for automatic CTR calculation to assist radiologists with rapid diagnosis of cardiomegaly and thus optimise the radiology flow.

Material and Methods: The study population included 1012 posteroanterior CXRs from a single institution. The Attention U-Net DL architecture was used for the automatic calculation of CTR. An observer performance test was conducted to assess the radiologist's performance in diagnosing cardiomegaly with and without artificial intelligence assistance.

Results: U-Net model exhibited a sensitivity of 0.80 [95% CI: 0.75, 0.85], specificity >99%, precision of 0.99 [95% CI: 0.98, 1], and a F1 score of 0.88 [95% CI: 0.85, 0.91]. Furthermore, the sensitivity of the reviewing radiologist in identifying cardiomegaly increased from 40.50% to 88.4% when aided by the AI-generated CTR.

Conclusion: Our segmentation-based AI model demonstrated high specificity (>99%) and sensitivity (80%) for CTR calculation. The performance of the radiologist on the observer performance test improved significantly with provision of AI assistance. A DL-based segmentation model for rapid quantification of CTR can therefore have significant potential to be used in clinical workflows by reducing radiologists' burden and alerting to an abnormal enlarged heart early on.

Keywords
cardiomegaly, deep learning, segmentation, chest radiograph, artificial learning

Introduction

Chest radiography is amongst the most commonly used diagnostic imaging techniques for disorders of the cardiothoracic and pulmonary systems. Chest radiographs account for 40% of the nearly 3.6 billion radiological investigations advised annually with the number of chest radiographs per 1000 people standing at 236 annually.[1] In fact, this is one of the most basic imaging investigations advised in nearly all

[1]Department of Radiodiagnosis, Dr DY Patil Medical College, Hospital and Research Center, DY Patil Vidyapeeth, DPU, Pune, India
[2]DeepTek Medical Imaging Pvt. Ltd, Pune, India
[3]Department of Radiology and Biomedical Imaging, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Corresponding author:
Pranav Ajmera, Department of Radiodiagnosis, Dr DY Patil Medical College, Hospital and Research Center, Dr D.Y. Patil Vidyapeeth, DPU, Pune 411018, India.
Email: pranavajmera@gmail.com

hospitalised patients, either as part of disease evaluation or as part of the usual workup for pre-operative assessment.

The currently used method for the assessment of CTR is manual; expert radiologists use manual segmentation to calculate the ratio. This is a time-consuming method. A CTR greater than 55% may indicate cardiomegaly, which is an indicator of multiple conditions, including hypertension, coronary artery disease, cardiomyopathies, and valvular heart disease. In fact, cardiomegaly is associated with adverse outcomes in patients with cardiac diseases and, therefore, may be an important clue to the initiation of early treatment.[2-4]Although traditionally CTR was described for detecting cardiac enlargement on chest radiographs, in the past decade research has shown the advantages cardiac MRI and that cardiac MRI is a much better modality to comment on actual cardiac enlargement. However, MRI as a modality is expensive, and far less accessible than a radiography setup and the number of cardiac MRI performed are far lesser than the number of chest radiographs performed each year. Cardiac MRI is performed only in selected patients who present to the hospital with cardiac symptoms and often if there is an indication detected on preliminary echocardiography findings. Therefore, it only follows that the chances of incidentally detecting cardiomegaly particularly in cases which are not symptomatic and are presenting for an unrelated disease and hence will not undergo a specialised cardiac MRI is much higher on a chest radiograph. This mandates the need for defining a cut-off for cardiac enlargement and while a CTR between 45 and 55% increases the sensitivity of detecting cardiomegaly, the trade-off is a lower specificity. Direct comparisons between chest radiographs and cardiac MRI yielded that while the discriminatory power of CTR with respect to MRI is weak, the agreement is higher when a cut-off of 55% is utilised. In effect, while no single CTR cut-off is completely sensitive and specific, utilising 55% as a cut-off balances the slightly lower sensitivity with a significantly higher specificity and hence a CTR greater than 55% should be considered highly indicative of true cardiomegaly and further diagnostically evaluated.[2]

Country-wise figures for the availability of radiologists as a percentage of the overall workforce show a declining trend. In Europe, there are 4–13 radiologists per 1,00,000 people with UK at the lower end of spectrum at 4.7 and Sweden at the higher end; in USA, the number stands at 10. The situation in developing countries such as Africa is much worse, with the ratio being under one radiologist per 1,00,000 population.[5,6] With an increase in the population and the number of radiological examinations performed each year, the relative number of qualified radiologists is declining, resulting in backlogs and delays in timely medical imaging even in large organisations like the UK-National Health Service and the US-Department of Veterans Affairs.[7-9] Consequent to the increase in the volume of scans, and a decrease in the number of radiologists, are the issues of increasing radiologist burn-out and high interpretation errors. Therefore, having in place a system which can automate the task of quantification on chest radiograph will imply one less thing for the radiologist to worry about while ensuring that this important parameter is not skipped over from radiology reports to conserve time.

Herein, we evaluated the performance of our previously developed Attention U-Net model that demonstrated excellent specificity in automatically calculating the CTR by testing it on an external validation dataset. To assess the actual clinical utility of employing this tool in a healthcare setting, we have subsequently performed an observer performance test with an experienced radiologist. Our segmentation model classified CXRs into two categories based on the CTR cut-off of 0.55: cardiomegaly present or cardiomegaly absent.

Material and Methods

Data collection

A total of 1257 sequential CXRs were acquired retrospectively between the period of January and March 2021. These chest radiographs were from a single institution and acquired on multiple machines of different milliamperage (mA). These included multiple computed radiography (CR) systems – SIEMENS 500 mA HELIOPHOS-D, SIEMENS 100 mA GENIUS-100R, SIEMENS 300 mA MULTI-PHOS-15R and a 600 mA digital radiography (DR) system, the SIEMENS MULTISELECT DR. For CR systems, AGFA 14x17 inch plate was used for adults (13–93 years). For the Digital radiography (DR) system, the SIEMENS 14x17 inch detector was used.

Out of the 1257 CXRs, 1012 CXRs were selected to be a part of the test set based on the inclusion and exclusion criteria. Chest X-rays which were acquired in an AP view, an oblique orientation, an expiratory position, or with too many motion artefacts created by large foreign bodies, overlying the cardiac contour were eliminated. Chest X-rays with pathological processes which obscured the cardiac contour like massive pleural effusion or dense consolidation completely obscuring the lung bordering cardiac contour were also excluded.

Study population

The study was approved by the Institutional Ethics Committee of our tertiary care hospital and research centre (DYPV/EC/642/2021), and the need for explicit written informed consent was waived. The study followed Health Information Portability and Accountability Act (HIPAA) standards for data management and anonymisation.

Amongst 1012 CXRs, the distribution of males and females was 61.56% (623 CXRs) and 38.44% (389 CXRs), respectively, with an age range of 13–93 years. The mean age of the cohort was nearly 42.6 years.

AI models and architecture

In our previous report by Gupte et al., 2021,[10] we compared three different U-Net based architectures on a hold-out dataset. The three multi-class segmentation models architectures we compared-first one (Model-1) was an enhanced UNET with spatial Attention Gate and Xception encoder,[11,12] the second one (model-2) was UNET with squeeze and excitation network blocks incorporated with ResNext-50 [13,14] and thirdly (model-3), a simple UNET with Efficient Net b4 encoder.[15,16] All of these models were as a first step pre-trained on an ImageNet weight. The training was performed on a dataset of 3416 CXRs. Both the training and validation dataset was the same for all three models. The Adam optimiser and binary cross-entropy loss were used with a learning rate that was reduced on the plateau, that is, the learning rate was reduced when the metric (validation loss) stopped improving. For each CXR, the model predicts two regions of interest for both the mediastinum and the chest. Although inferencing, to obtain a clear-cut mask, the predicted pixel probabilities were thresholded. The pixel probabilities less than the threshold value were converted to 0, while those greater than the threshold value were converted to 1. The output masks were subjected to morphological transformations to reduce noise and error. The mask, of size 512x512 pixels, was eroded for 2 iterations with a kernel of 3x3 and then dilated for 1 iteration with a kernel of size 3x3. This removed all the noisy pixels on the outer edges giving a clean mask. Although the difference in the output masks may not be apparent to human eyes, we observe MAE reduction by 12.5% and RMSE by 9%. Bounding boxes were constructed using the extreme pixels of the segmented mask generated by the model for the heart and the chest. The coordinates of the bounding boxes were used to calculate the CTR by determining the width of the chest and heart.

All of these models were trained on a common dataset to determine which one of them outperformed the other two and was most robust. Although primarily all were UNET based architectures, the encoders and the intermediate layers were different for each model. To ensure better training, we performed hyperparameter optimisation, image augmentation, and image processing. These were tested on a hold-out test set of 183 CXRs to evaluate model performance. The performance of all the three models was close: Model-1 had a sensitivity of 0.96, specificity of 0.81; Model-2 had a sensitivity of 0.87 and specificity of 0.86; Model-3 had a sensitivity of 0.94 and specificity of 0.83. A comparison of the f-1score revealed the model-3 (0.88) to marginally outperform both model-2 (0.86) and model-1 (0.87).

Although it was not possible to choose the best performing model based only on sensitivity, specificity and f-1 score. A combined comparison based on the regression metrics-mean absolute error (MAE) and root mean squared error (RMSE) revealed that model-1 (MAE = 0.0209; RMSE= 0.0312) slightly outperformed the other two, model-2 (MAE= 0.0206; RMSE= 0.0317), model-3 (MAE= 0.0328; RMSE= 0.0798). Based on the performance, we selected the Attention U-Net model for external validation on the institutional dataset. Figure 1 is a representation of the steps involved in processing of a chest radiograph by the AI model to quantify it.

Establishing ground truth

To establish the ground truth, a consensus opinion of three Board certified radiologists with 5 (Radiologist-A1), 7 (Radiologist- A2) and 23 (Radiologist-A3) years of experience was undertaken. Radiologist A2 had received dedicated 1-year of subspeciality training in cross-sectional imaging while Radiologist A3 had thoracic subspeciality training and 23 years of experience. They were tasked with manually calculating and annotating the CTR and classifying the CXR as cardiomegaly positive or negative. The cardiac size was measured by drawing a straight line down the most lateral points of the cardiac margins and calculating the diameter between them. To measure the thoracic border, a straight line from the inner margin of the rib cage on one side to the opposite was drawn. Out of the 1012 CXRs, 242 CXRs had a CTR greater than or equal to 0.55 (positive for cardiomegaly) and 770 CXRs had a CTR less than 0.55. The combined output of the three radiologists is hereafter, referred to as the ground truth (Radiologist-A)

The observer performance test

The observer performance test was conducted to compare the performance of the model with that of the radiologist-A (ground truth) and to determine whether the performance of another radiologist (Radiologist-B) be improved with the aid of the model. The radiologist-A who participated in the establishment of the ground truth were excluded from the observer performance test. The test proceeded in two phases. In the first phase, the CTR calculated by the model was concealed, and the observer (radiologist-B) independently assessed each CXR for the presence or absence of cardiomegaly without the help of the model. He was explained in brief about the purpose of categorising for cardiomegaly and was instructed to follow the method he uses

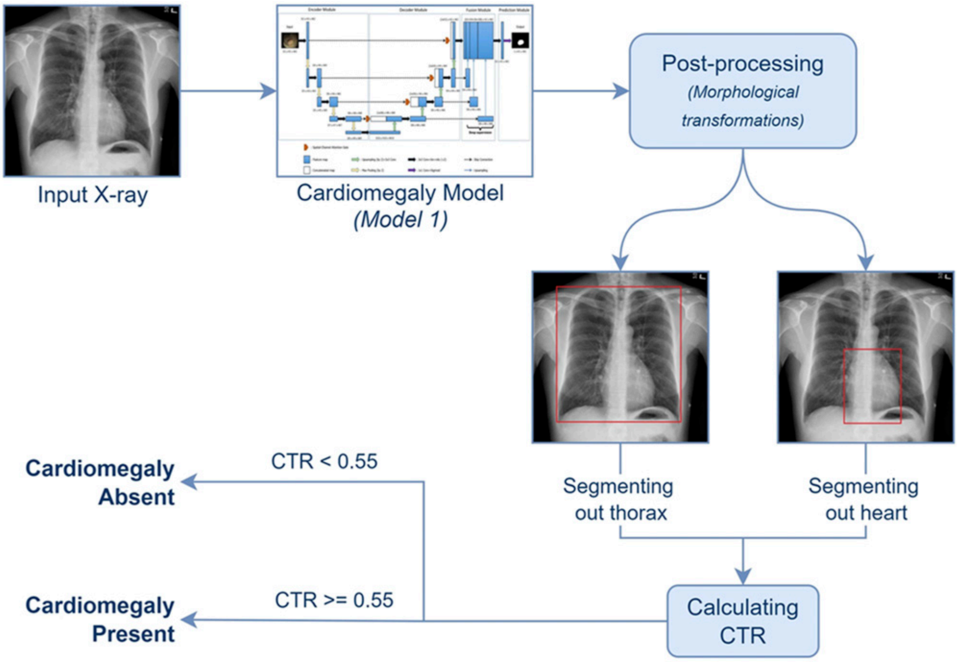

Figure 1. The flowchart depicting the process involved in the processing and post-processing of the raw data (chest radiograph) to quantify the CTR and categorise the result as positive or negative for cardiomegaly.

to categorise in day-to-day reporting, which was the visual method of assessment. In the second phase, after a gap of one month to avoid memory bias, the observer re-evaluated each CXR with the aid of the model. Finally, the results of both the phases were compared to measure the agreement on the detection of cardiomegaly.

Statistical analysis

The performance of the annotating radiologist and the AI model was evaluated using classification metrics (sensitivity, specificity, precision and F1 score) and regression metrics (Mean Absolute Error and Mean Squared Error). The coefficient of determination (r squared value) was calculated to determine the goodness of fit of the model for the given dataset. Paired t-test was used to compare the mean difference between CTR calculated by the model and CTR annotated by the radiologist. Intra class-correlation statistical analysis was used to calculate the degree of agreement between the two continuous variables (CTR for each individual chest radiograph) – AI model calculated and the radiologist-A annotated CTR. However, Bland–Altman analysis was utilised to interpret the agreement based on the mean difference between the two variables.

Specificity, sensitivity, NPV, PPV, accuracy and kappa score were used to compare the performance of the model with that of the reporting radiologist. These metrics were also used to compare the performance of the aided versus unaided radiologist.

Results

Diagnostic performance of CTR calculation model

The diagnostic performance of the CTR calculation model (Model-1) is summarised in Table 1. The classification metrics demonstrate how well the model performs in clinical settings. The model was highly specific in determining the CTR, with a specificity nearly 1.00 [95% CI: 0.99,1]. The model achieved a sensitivity of 0.80 [95% CI: 0.75, 0.85], precision of 0.99 [95% CI: 0.98, 1.0], and a F1 score of 0.88 [95% CI: 0.85, 0.91]. The reliability of the CTR calculation was determined by evaluating regression metrics. The model performed extremely well in calculating CTR with a MAE of 0.0254 ± 0.06 and an MSE of 0.0016 ± 0.014. We also computed the confusion matrix for categorising patients in the CTR ranging less than or greater than 0.55. Most of the CTR values annotated by the radiologist and calculated by the model were found to be in great agreement. Out of 1012 CXRs, 768 CXRs that were annotated as CTR<0.55 by the radiologist were also predicted the same by the model. Additionally, 193 CXRs that were annotated as CTR≥0.55 by the radiologist were also predicted the same by the model (Table 2). During the

Table 1. Classification and regression metrics to evaluate the performance of the AI model-1.

	Metrics	Value
Classification metrics [95% CI]	Sensitivity	0.80 [0.75, 0.85]
	Specificity	1.00 [0.99, 1.0]
	Precision	0.99 [0.98, 1.0]
	F-1 score	0.88 [0.85, 0.91]
Regression metrics (± SD.)	Mean absolute error (MAE)	0.0254 ± 0.06
	Mean squared error (MSE)	0.0016 ± 0.014

Table 2. Confusion Matrix of Attention U- Net model for CXR image dataset.

	Predicted (CTR <0.55)	Predicted (CTR ≥0.55)
Annotated (CTR <0.55)	768	2
Annotated (CTR ≥0.55)	49	193

analysis, we discovered that 636 out of 1012 CXRs, that is 62.84% of all predictions, were within +/− 5% error.

Performance of the Deep Learning Model versus Ground truth

The performance of the model as compared to that of the board-certified radiologist (ground truth) can be visualised in the scatter plot (Figure 2). The coefficient of determination (r squared value) for the same was observed to be 0.809, indicating that the regression model fits the observed data with low variance. The scatter plot also indicates that out of 1012 samples, only 51 samples were misclassified. There was one particular radiograph with a large discrepancy between the ground truth annotated (0.5) and predicted value (0.2), the cause for this was the error in proper detection of the left cardiac border, which resulted in an erroneously small reading of the cardiac border.

We performed a paired t-test between the two continuous variables of CTR calculations on the DL model and the ground truth. The analysis revealed that the mean values of CTR calculation by the DL model was slightly lower with a minimal difference of 0.0079764, which was statistically significant with a p-value of <0.001 (Supplemental Table 1) across the entire dataset. However, since the difference was minimal, the actual impact on classification of the model was negligible. Additionally, an intra-class correlation coefficient of the two variables revealed an excellent agreement between the two continuous variables to measure CTR (Supplemental Table 2). Therefore, while the DL model consistently underpredicted CTR when compared to the ground truth, the difference was minimal and there was an excellent correlation between ground truth and the DL model predicted CTR values.

Bland–Altman plot of the difference between the CTR calculations of the ground truth and those calculated by the DL model reveal a high agreement between the two (Figure 3).

AI outcome

The segmentation-based Attention U-Net model predicted the cardiothoracic ratio with high specificity and sensitivity. The results obtained by using this approach corresponded to the annotations made by the expert radiologist (Figure 4). Although the model was highly specific in calculating the CTR, there were some instances where the model misclassified the cases. A sample X-ray image of misclassification is shown in Figure 5. As shown in the figure, our model predicted the borderline cardiomegaly condition with a CTR=0.53, while the radiologist annotated it with a CTR=0.56. Although there was a small difference between the CTR values predicted by the model and those annotated by the radiologist, these misclassifications were unavoidable due to observer and method variations in the dataset. Additionally, the prediction made by the model can give a second chance to the radiologist to review the scan before making the final decision.

Observer performance test (Aided vs Unaided experiment):

To assess the clinical utility of the model and its impact on the performance of the reporting radiologist, we conducted an experiment whereby all 1012 chest radiographs were assessed for the presence or absence of cardiomegaly by a radiologist (radiologist-B) with 5-years of experience. The entire experiment was conducted in two phases: in the first

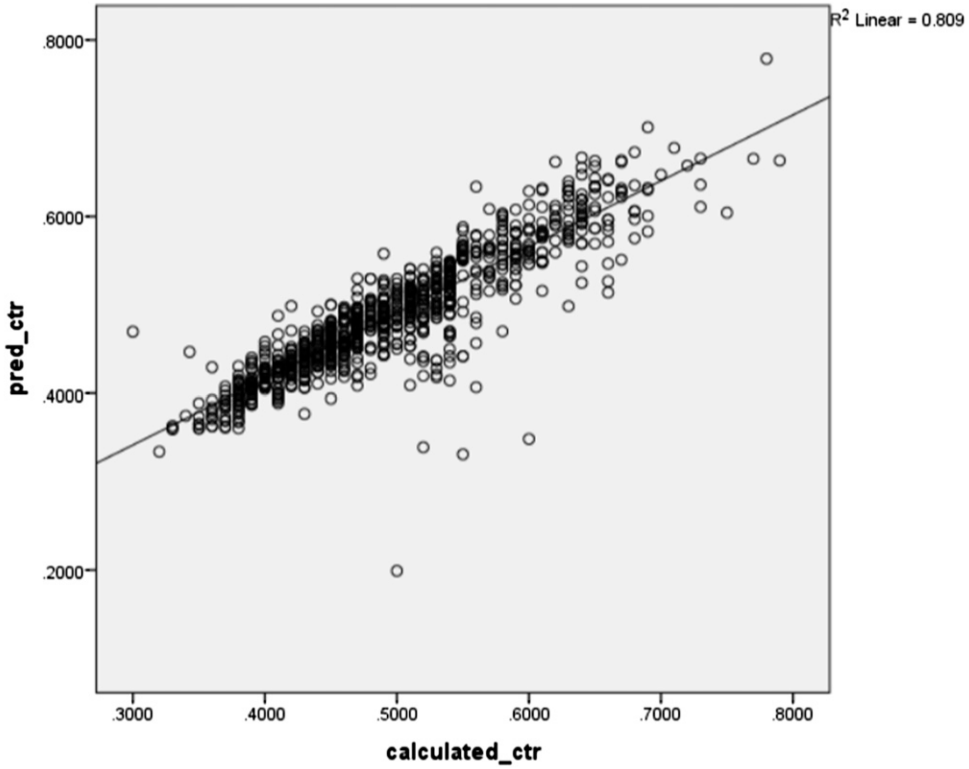

Figure 2. The scatter plot between calculated (radiologist's annotated/ground truth) CTR (X-axis) versus predicted (model-calculated) CTR (Y-axis) indicates a close agreement between the two variables. A closer analysis of the plot reveals 51 misclassifications with agreement on all of the remaining 961 cases.

phase, radiologist-B was not alerted to the pathology and was simply asked to annotate all 1012 chest radiographs. For this purpose, the radiologist was not provided with the AI aid to read the radiograph ('Unaided'). We compared the performance of the unaided radiologist in classifying a CXR as positive for cardiomegaly with the actual cases of cardiomegaly based on the ground truth.

The analysis revealed that the unaided radiologist had a sensitivity of 40.5% and specificity of 99.9% in detecting cardiomegaly. The radiologist was in agreement with the gold standard (CTR>0.55 as cardiomegaly) on 867 out of 1012 cases and demonstrated a positive predictive value (PPV) of 99%. Negative predictive value (NPV) of 84.2% and diagnostic accuracy of 85.67% (Table 3). The Kappa value of 0.506 indicates a moderate agreement, with a *p*-value of <0.001 (Supplemental Table 3).

For the second part of the experiment ('Aided'), the same radiologist re-assessed the entire dataset for multi-class pathologies with the aid of the AI model (the model created bounding boxes on the chest radiograph with pre-calculated CTR for each radiograph). To eliminate memory bias, the aided experiment was conducted one month after the unaided experiment.

The analysis revealed that the aided radiologist had a sensitivity of 88.4%, specificity of 89.9%, the PPV of 73.3% and NPV of 96.1%. The aided radiologist was in agreement with the gold standard on 906 out of 1012 cases and had a diagnostic accuracy of 89.53% (Table 3). Also, the individual diagnostic accuracy of the DL model was far superior with a value of nearly 94.96%. The Kappa value of 0.731 indicates very good agreement with a *p*-value of <0.001 (Supplemental Table 4). The overall f-1 score also improved from 0.57 for the unaided radiologist to 0.83 for the aided radiologist. Table 3 compares different parameters for the diagnostic performance of the model, unaided radiologist and unaided radiologist. The tabulated results clearly indicate that the sensitivity of the radiologist increased from 40.5% to 88.40% with AI assistance. There was also an improvement in the negative predictive value (from 85.67% to 89.53%) and the overall diagnostic accuracy (85.67%–89.53%) of the aided radiologist when compared to the unaided radiologist. Although the overall performance of the radiologist improved when assisted by the AI system, the stand-alone performance of the DL model was still statistically better.

Time factor

The stand-alone DL based model processed a radiograph in approximately less than 2s in all cases. Although manual

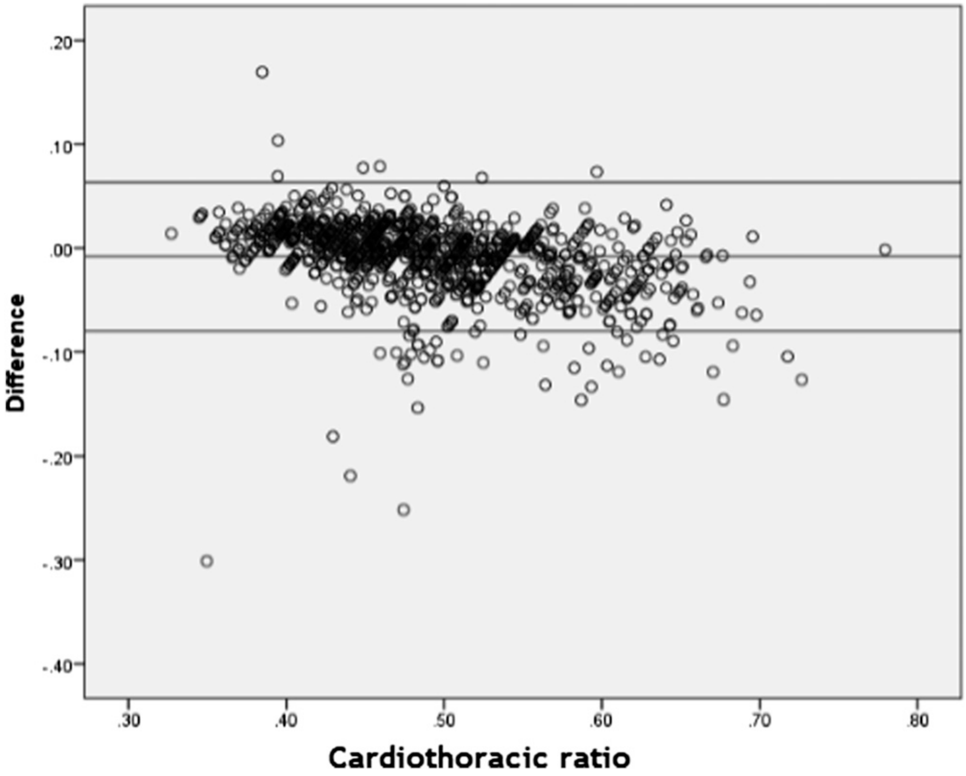

Figure 3. Bland–Altman plot to assess the mean difference in calculation of CTR by the ground truth (radiologist annotated) and the DL model.

measurements of the CTR required approximately 19.5 ± 3s, the AI assisted radiologist could make the same CTR measurements in 4.2 ± 1.8s (factoring in both the time taken to generate the AI annotation and the radiologist to either agree with the same or reposition the boxes for a more accurate fit. Thus, the AI assisted method of quantification was a little over 4 times faster than the manual method.

Discussion

Cardiothoracic ratio (CTR) obtained from CXR is an important parameter for assessing heart diseases, particularly cardiomegaly.[17,18] However, measuring it necessitates manual measurements that are time-intensive and user-dependent. Despite its utility and merits, the measurement of CTR is burdensome in clinical practice. Additionally, manual calculation of CTR introduces subjectivity into the diagnosis, and many borderline cases may go undetected or incorrectly diagnosed. Recently, the utility of AI-based tools in calculating the CTR has been technically corroborated in many studies.[19-21] In our previous report (xxx, 2021),[10] we compared three different segmentation model architectures for the calculation of CTR and observed that Attention U-Net yielded better results than EfficientNet U-Net and SE-Resnext U-Net. In the current study, we used the Attention U-Net deep-learning model to calculate the CTR in a clinical setting. We observed that the CTR values derived from the Attention U-Net model exhibited excellent agreement with the CTR values annotated by the radiologist. Out of 1012 samples, only 51 samples were misclassified by the model, and that too by a small margin, while 961 samples were completely consistent with the annotations provided by the radiologist. It should be noted that most of the misclassifications were for the borderline conditions, with CTRs near 0.55 or chest radiographs with moderate pleural effusion[24] which was partially obscuring the thoracic contour of lung and thus generating an erroneously larger CTR. Although the misclassification of cardiomegaly is unavoidable when the CTR is close to 0.55, this is important because, in case CTR is more than 0.55, it can alert the concerned radiologist and allow the radiologist to review the scan again before making a final decision. The pragmatic approach of using model-generated CTR in assisting the radiologist proved to be beneficial in improving the diagnosis of cardiomegaly when compared to the unaided radiologist. Since borderline cases are frequently missed, the practicability of the model will help in the radiologist's objective decision making. Erroneous results are a problem faced by all DL

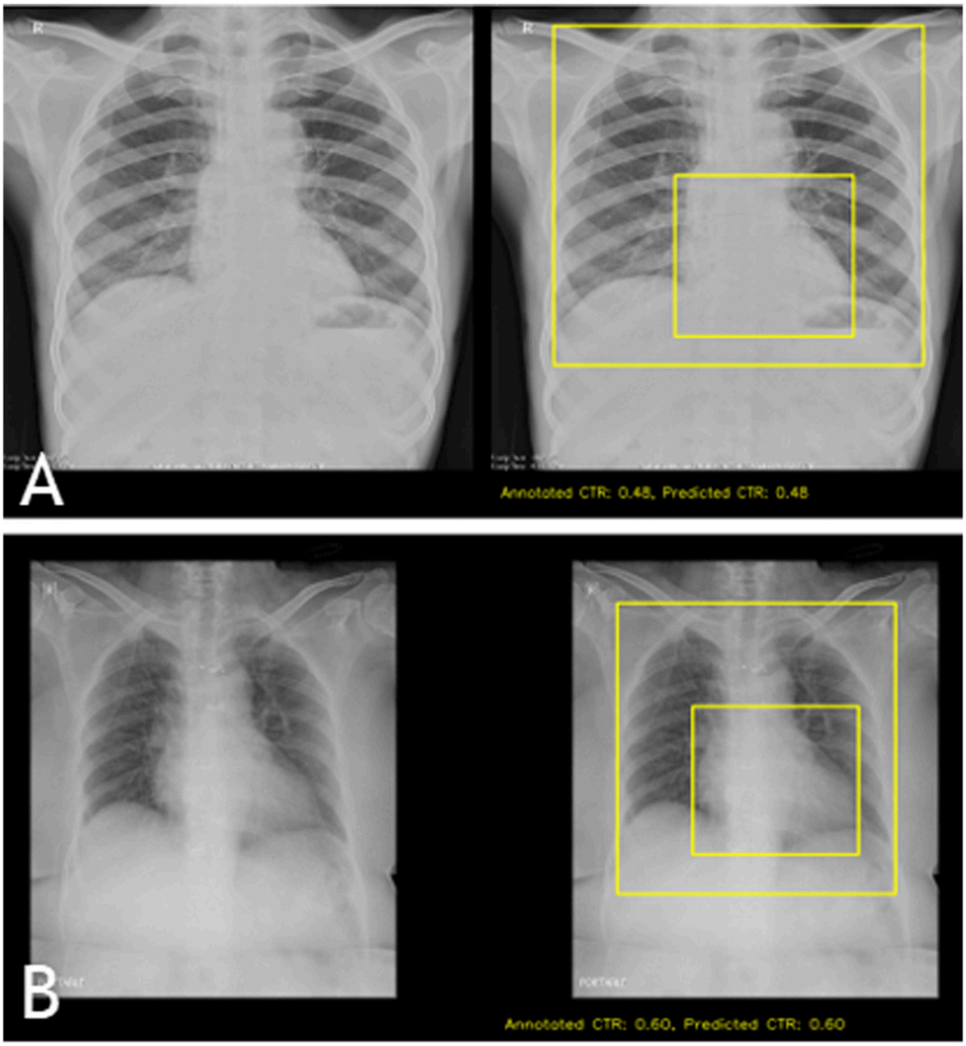
R
A
Annotated CTR: 0.48, Predicted CTR: 0.48
B
Annotated CTR: 0.60, Predicted CTR: 0.60

Figure 4. CXRs representing the Cardiothoracic ratio annotated by the radiologist-A (ground truth) and predicted by the model. Bounding boxes correspond to the segmentation of the heart and thorax by the model. The predicted and the annotated values for CTR (A) less than 0.55 (Normal) and (B) greater than 0.55 (cardiomegaly) were in complete agreement with each other.

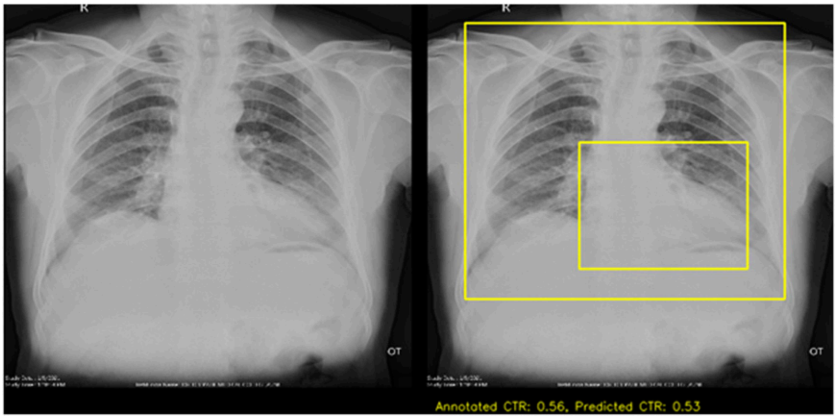

Figure 5. An example of misclassification. Bounding boxes represent the segmentation of the heart and thorax by the model. In this case, the radiologist annotated the CTR as 0.56, while the model predicted the CTR as 0.53.

Table 3. Comparison between the diagnostic performance of the AI model, unaided and aided radiologist.

Parameter	True negative i.e calculated CTR (ground truth) cutoff 0.55: No cardiomegaly (<0.55) & test: Negative	True positive i.e calculated CTR (ground truth) cutoff 0.55: Cardiomegaly (≥0.55) & test: Positive	False negative i.e calculated CTR (ground truth) cutoff 0.55: Cardiomegaly (≥0.55) & test: Negative	False positive i.e calculated CTR (ground truth) cutoff 0.55: No cardiomegaly (<0.55) & test: Positive	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Diagnostic accuracy, %	Gold standard
Model-I	768	193	49	2	79.80	99.70	99.00	94.00	94.96	Ground truth
RadiologistUnaided	769	98	144	1	40.50	99.90	99.00	84.20	85.67	Ground truth
RadiologistAided	692	214	28	78	88.40	89.90	73.30	96.10	89.53	Ground truth

models in scenarios where the contours are not clearly demarcated, as such cardiac contour has a more major impact than the thoracic contour. [22]

Saiviroonporn et al.[23] used the VGG-16 U-Net model to assess the impact of AI-enabled reporting of CTR. Their findings concluded that AI alone had higher variations than human readers, but it could support the radiologist by reducing observer variation and operation time.[24] Another study published by Li et al.[20] for the calculation of CTR using a custom 2D-U-Net architecture discovered a good match between the performance of the DL model and the human reader, with the difference statistically insignificant. Chamveha et al.[24] used U-Net architecture with the VGG-16 model and demonstrated that nearly 76.5% of the CTR measurements performed by the DL model were acceptable to the human reader, resulting in significant time savings.

Our model correctly classified cardiomegaly in 94.9% of the cases, and 63.8% of all CTR calculations were within ±5% error of that calculated by the expert reader. The Attention U-Net based deep learning model achieved an excellent specificity of 100% and a precision of 99% on the clinical test dataset, which has never been reported before to our knowledge. Additionally, the radiologist aided by AI showed relatively higher sensitivity and negative predictive value (88.4% and 96.10%, respectively) as compared to the unaided radiologist (40.50% and 84.20%, respectively). Although the number of false positives for aided radiologist was much higher at 78, the number of false negatives was significantly lower at 28. Curiously the number of false negatives and positives for the DL model stand-alone was much lower. Despite the increase in FP by the aided radiologist, the probability of the radiologist missing out on the cases of true cardiomegaly which potentially can in future lead to increased cardiac problems was much lower (FN)

Most of the scans (37 out of 39) with cardiomegaly, which were missed by the radiologist-B assessing cardiomegaly visually without AI aid, were in the borderline category (CTR=0.55–0.60), and after the AI aid, these were then correctly classified by the radiologists as cardiomegaly. As a result, we conclude that our DL model based on U-Net architecture and segmentation is successful in accurately calculating CTR. An improvement in the radiologist's performance, when aided by the model, illustrates the model's reliability and utility. The incorporation of the AI model into the radiology workflow for the detection of cardiomegaly can save a radiologist's critical time and allow the reader to allocate more time to look for subtle and suspicious pathologies. The diagnostic accuracy of the AI model alone was greater than that of the aided radiologist, which could be attributed to the borderline cases that were misclassified by the radiologist.

Our approach to testing the model had certain limitations. Although our model proved to be highly specific and precise in calculating the CTR, the test set included data from a single hospital. Future studies with data from multiple institutions/hospitals will be beneficial for increasing the generalizability and robustness of the model. The DL model for CTR calculation demonstrated excellent performance in the retrospective analysis; whether the model can be implemented prospectively in clinical practice can be validated in future research. It is this prospective deployment in clinical setting which will establish the actual clinical utility, as then the model will be exposed to certain factors unique to the real world – like difference in the image due to inspiratory/expiratory position. With increasing deployment of ID based categorisation of patient investigations into a single digital file, an opportunity exists for the DL model to draw inferences from multiple radiographs to comment on the change of CTR in a patient over time. However, this will need to be evaluated in the real world with specific training of the model to detect, infer and represent the radiograph comparisons.

In conclusion, our research presents a simple and concise approach for calculating CTR from chest radiographs. Our model achieved a sensitivity of 80% for a specificity over 99% in calculating the CTR. The observer performance test demonstrated a significant statistical improvement in the performance of the clinician when AI assistance was provided and also a multi-fold decrement in time spent on quantifying CTR. Apart from reducing the likelihood of missed finding due to inter-observer variations and saving radiologists' effort by avoiding manual calculation of CTR, our approach has the added benefit of alerting radiologists to the borderline cases that are not evident to the unaided human eye and thus providing the algorithmic second opinion. Since, an AI-alone system can occasionally misclassify the condition, a human-in-the-loop approach can leverage both humans and AI to repeatedly perform the task accurately.

Author contributions

PA was involved in conceptualisation, methodology, validation, formal analysis, investigation, data curation, writing and visualisation. AK was involved in conceptualisation, methodology, project administration, resources, data curation, writing and visualisation. TG was involved in technical writing, software and formal analysis. RP was involved in writing and analysis. VK was involved in data curation, technical writing and project administration. VD was involved in writing, validation and visualisation. All the authors made significant contributions to this project and the study was approved by all of them.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: A.K. is a Professor at Dr DY Patil Medical College, Pune and is also a co-founder of DeepTek Inc., whose expertise was employed to build this model. V.D. is a Professor of Radiology at Keck school of Medicine, USC, USA and is also on the advisory board of DeepTek Inc; he is a consultant to Radmetrix Inc, Cohere Inc and Westat Inc.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethics approval

The study was approved by the Institutional Ethics Committee of our tertiary care hospital and research centre (Dr D.Y. Patil Vidyapeeth) in its ethical committee meeting as DYPV/EC/642/2021. Ethical clearance granted by the Institutional Review Board, as mentioned in the manuscript.

Informed Consent

The need for explicit written informed consent was waived.

Data availability

The datasets generated and analysed during the current study are not publicly available to ensure complete patient data anonymity but are available from the corresponding author on reasonable request.

Disclaimer

The views expressed in the submitted article are solely the authors and are not an official position of the Institution.

ORCID iDs

Pranav Ajmera https://orcid.org/0000-0001-8801-0235
Vinay Duddalwar https://orcid.org/0000-0002-4808-5715

Supplemental Material

Supplemental material for this article is available online.

References

1. Speets AM, van der Graaf Y, Hoes AW, et al. Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. Br Journal General Practice : The Journal R Coll Gen Pract 2006; 56: 574–578.
2. Simkus P, Gutierrez Gimeno M, Banisauskaite A, et al. Limitations of cardiothoracic ratio derived from chest radiographs to predict real heart size: comparison with magnetic resonance imaging. Insights Into Imaging 2021; 12: 158–160.
3. Philbin E. F., Garg R, Danisa K, et al. The relationship between cardiothoracic ratio and left ventricular ejection fraction in congestive heart failure. Arch Intern Med 1998; 158: 501–506.

4. Brakohiapa EKK, Botwe BO, Sarkodie BD, et al. Radiographic determination of cardiomegaly using cardiothoracic ratio and transverse cardiac diameter: Can one size fit all? Part one. Pan Afr Medical Journal 2017; 27: 201.
5. https://www.telemedicineclinic.com/wp-content/uploads/2016/11/Europes_looming_radiology_capacity_challenge-A_comparitive_study.pdf (Accessed: 20th November,2021).
6. Iyawe EP, Idowu BM, Omoleye OJ. Radiology subspecialisation in Africa: A review of the current status. SA J Radiol 2021; 25: 1–7.
7. Çallı E, Sogancioglu E, van Ginneken B, et al. Deep Learning for Chest X-ray Analysis: A Survey. Med Image Anal 2021; 72: 102125.
8. Ropp A, Waite S, Reede D, et al. Did I miss that: subtle and commonly missed findings on chest radiographs. Curr Probl Diagn Radiol 2015; 44: 277–289.
9. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ 2017; 359: j4683.
10. Gupte T, Niljikar M, Gawali M, Kulkarni V, Kharat A, Pant A. Deep learning models for calculation of cardiothoracic ratio from chest radiographs for assisted diagnosis of cardiomegaly. In: 2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), 5 August 2021 (pp. 1–6). IEEE.
11. Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint 2018 :1804.03999.
12. Khanh TLB, Dao D-P, Ho N-H, et al. Enhancing u-net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. Appl Sci 2020; 10: 5729.
13. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. IEEE/CVF Conf Comp Vis Pattern Recognition 2018; 2018: 7132–7141.
14. Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: InProceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017, 1492–1500.
15. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. InInternational Conf Machine Learn 2019; 24: 6105–6114. PMLR.
16. Baheti B, Innani S, Gajre S, et al. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. InProceedings IEEE/CVF Conf Comp Vis Pattern Recognition Workshops 2020; 2016: 358–359.
17. Dimopoulos K, Giannakoulas G, Bendayan I, et al. Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. Int J Cardiol 2013; 166: 453–457.
18. Okute Y, Shoji T, Hayashi T, et al. Cardiothoracic Ratio as a Predictor of Cardiovascular Events in a Cohort of Hemodialysis Patients. J Atheroscler Thromb 2017; 24: 412–421.
19. Arsalan M, Owais M, Mahmood T, et al. Artificial Intelligence-Based Diagnosis of Cardiac and Related Diseases. J Clinical Medicine 2020; 9(9): 871.
20. Li Z, Hou Z, Chen C, et al. Automatic cardiothoracic ratio calculation with deep learning. IEEE Access 2019; 7: 37749–37756.
21. Que Q, Tang Z, Wang R, et al. CardioXNet: Automated Detection for Cardiomegaly Based on Deep Learning. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf 2018; 2018: 612–615.
22. Lee MS, Kim YS, Kim M, et al. Evaluation of the feasibility of explainable computer-aided detection of cardiomegaly on chest radiographs using deep learning. Scientific Rep 2021; 11: 16885.
23. Saiviroonporn P, Rodbangyang K, Tongdee T, et al. Cardiothoracic ratio measurement using artificial intelligence: observer and method validation studies. BMC Med Imaging 2021; 21: 1–11.
24. Chamveha I, Promwiset T, Tongdee T, et al. Automated Cardiothoracic Ratio Calculation and Cardiomegaly Detection Using Deep Learning Approach, 2002. arXiv preprint 2020.07468.