Multimodal Cardiomegaly Classification with Image-Derived Digital Biomarkers

Benjamin Duvieusart[1], Felix Krones[2], Guy Parsons[3], Lionel Tarassenko[1],
Bartłomiej W. Papież[4,5], and Adam Mahdi[2](✉)

[1] Department of Engineering Science, University of Oxford, Oxford, UK
[2] Oxford Internet Institute, University of Oxford, Oxford, UK
adam.mahdi@oii.ox.ac.uk
[3] Intensive Care Registrar, Thames Valley Deanery, NIHR Academic Clinical Fellow at Oxford University, Oxford, UK
[4] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[5] Nuffield Department of Population Health, University of Oxford, Oxford, UK

Abstract. We investigate the problem of automatic cardiomegaly diagnosis. We approach this by developing classifiers using multimodal data enhanced by two image-derived digital biomarkers, the cardiothoracic ratio (CTR) and the cardiopulmonary area ratio (CPAR). The CTR and CPAR values are estimated using segmentation and detection models. These are then integrated into a multimodal network trained simultaneously on chest radiographs and ICU data (vital sign values, laboratory values and metadata). We compare the predictive power of different data configurations with and without the digital biomarkers. There was a negligible performance difference between the XGBoost model containing only CTR and CPAR (accuracy 81.4%, F1 0.859, AUC 0.810) and black-box models which included full images (ResNet-50: accuracy 81.9%, F1 0.874, AUC 0.767; Multimodal: 81.9%, F1 0.873, AUC 0.768). We concluded that models incorporating domain knowledge-based digital biomarkers CTR and CPAR provide comparable performance to black-box multimodal approaches with the former providing better clinical explainability.

Keywords: Cardiomegaly · Multimodal approach · Domain knowledge · Digital biomarkers · Data fusion · Deep learning · Chest X-ray · Cardiothoracic ratio · Segmentation · Detection

1 Introduction

There is a worldwide shortage of trained radiologists [27,28]. The application of automated radiograph labelling and diagnosis algorithms to support clinical staff has the potential to increase the efficiency of clinical workflows and reduce demand on radiology services. A tool which accurately identifies pathology, as part of an appropriate care pathway, has the potential to increase the quality

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 13–27, 2022.
https://doi.org/10.1007/978-3-031-12053-4_2

of care worldwide. This paper approaches the problem of creating an automatic labelling tool for cardiomegaly.

Cardiomegaly, an abnormal enlargement of the heart, may result from many cardiac conditions, such as coronary artery disease or congenital heart disorders. Often cardiomegaly is first identified by examining a patient's cardiothoracic ratio (CTR), calculated by taking the ratio of the cardiac width to the thoracic width on a posterior-anterior projection of a chest X-ray. The cardiac width is measured as the horizontal distance between the leftmost and rightmost extremes of the cardiac shadow, and the thoracic width is measured as the horizontal distance from the inner margin of the ribs at the level of the hemidiaphragm. A CTR of 0.5 is usually classed as the upper limit for normal cardiac size and hence commonly used as the delimiter of cardiomegaly.

The increased availability of large publicly-available clinical imaging datasets has accelerated the use of computer vision and machine learning techniques to identify the CTR by using edge detection [14] or convolutional neural networks [25,33]. While CTR is an important and widely accepted first metric to identify cardiomegaly, there are inherent limitations. CTR values are prone to inaccuracies as both the cardiac and thoracic widths are dependant on many factors, such as the dilation of cardiac chambers, respiratory phase and body posture. It is known that this method risks flagging false-positives, causing many patients with suspected cardiomegaly to be subjected to further imaging. Despite these concerns, the CTR is still a fundamental tool for identifying cardiomegaly due to its simplicity and since false-positives are considered a more acceptable error type in clinical settings.

A clinician can compensate for the uncertainties associated with using only CTR values by synthesising all available patient data from multiple modalities, including patient metadata, vital signs and blood test results to refine a diagnosis of cardiomegaly and identify the underlying pathology. Multimodal approaches in machine learning have been tested to various degrees, such as through combining medical images with basic demographics to predict endovascular treatment outcomes [30], or classifying skin lesions using a combination of dermoscopic images and patient age and sex [7]. There have also been efforts to use multimodal data to classify cardiomegaly by combining imaging data, with extensive non-imaging data (patient metadata, lab results, and vital signs) [10].

In this paper, we consider the classification of cardiomegaly by mimicking existing diagnostic pathways - we combine domain knowledge in the form of two image-derived digital biomarkers with imaging and non-imaging data from the Intensive Care Unit (ICU). The digital biomarkers used here are CTR and the cardiopulmonary area ratio (CPAR), the latter being a proxy for the cardiothoracic area ratio which has been used to evaluate cardiac function [19]. While CTR is the classic clinical value, it only measures horizontal expansion, while the CPAR provides a more holistic measure of cardiac enlargement. We assess the predictive power of models using different combinations of data modalities: imaging data, non-imaging ICU data and combination of imaging and non-imaging data. Finally, we compare models incorporating domain knowledge-based digital biomarkers CTR and CPAR with the black-box multimodal approaches.

2 Data and Methods

2.1 Data Sources

We used four publicly available databases: MIMIC-CXR [9,17], MIMIC-IV [9,16], JSRT [31] and Montgomery County [2,15].

MIMIC-IV Database. This database contains medical data for 382,278 patients from the Beth Israel Deaconess Medical Center Intensive Care Units between 2008 and 2019. MIMIC-IV is structured into three sections: *core* (patient metadata, ward transfers), *icu* (vital sign time series, ICU procedures), and *hosp* (laboratory results, prescriptions).

MIMIC-CXR Database. This is a large publicly available database which contains 227,835 studies for 65,379 patients (a subset of the MIMIC-IV patients) from 2011 to 2016 collected from the Beth Israel Deaconess Medical Center Emergency Department. Each study contains one or more chest radiographs taken from different positions for a total of 377,110 images. Additionally, each study is accompanied by a semi-structured free-text radiology report describing the the findings of the radiologist. In this study we primarily use MIMIC-CXR-JPG [18] which is derived from MIMIC-CXR, containing the same images in the JPG format instead of the original DICOM format. While there is a certain loss of information by using JPG, the DICOM format can be difficult to use and comprehend, hence JPG format is preferred.

JSRT Database. The Japanese Society of Radiological Technology (JSRT) database is a publicly available database of posterior-anterior chest radiographs collected from medical centers in Japan and the USA. The database consists of 247 chest radiographs. The associated database, Segmentation Chest Radiographs [8], provides segmentation masks of lungs and heart.

Montgomery County Database. This is a publicly available database of chest radiographs collected from the Tuberculosis control program by the Department of Health and Human Services of Montgomery County, USA. It contains 138 chest radiographs, and contains segmentation masks of lungs.

2.2 Dataset Preparation

We prepared two new datasets, described below, one to train and test digital biomarkers models and the second to train and test the cardiomegaly classifiers.

CTR Dataset. The CTR dataset was created to train and test the models used to calculate the image-based digital biomarkers CTR and CPAR. It combines the JSRT, Montgomery County and MIMIC-CXR databases containing a total of 585 chest radiographs (247 from JSRT, 138 from Montgomery County, and 200 from MIMIC-CXR), and their associated segmentation masks for the heart and lungs. The JSRT database has an associated database, Segmentation Chest Radiographs, which contains segmentations of the heart and lungs. For Montgomery County Database, we used the included lung segmentations and supplemented this with manual, clinician supervised, segmentations of the heart. For MIMIC-CXR, we selected 200 random posterior-anterior chest radiographs with labels *fracture*, *consolidation* and *support devices*. Any samples also present in the cardiomegaly dataset, described below, were removed. Lung and heart segmentation masks for MIMIC-CXR images were manually completed under the supervision of a clinician. The manually completed segmentations will be released in due course.

Cardiomegaly Dataset. The cardiomegaly dataset was used to train and test the cardiomegaly classifiers. It combines data from MIMIC-CXR-JPG and MIMIC-IV. MIMIC-CXR-JPG comes with four cardiomegaly labels: *positive*, *negative*, *uncertain* and *no mention*. These labels were extracted from two natural language processing tools, NegBio [23] and CheXpert [13]. We only used images where both tools agreed on the label and further removed all *uncertain* and *no mention* labels, since in the last case we could not exclude cardiomegaly. This criteria reduced the size of the MIMIC-CXR dataset to 54,954 studies (81,346 radiographs) for 23,303 patients.

Cardiomegaly is identified from posterior-anterior radiographs, to avoid the unnatural enlargement of the cardiac silhouette which may occur from the anterior-posterior view. As such, we linked ICU stays from MIMIC-IV with the closest radiographic study containing a posterior-anterior chest radiograph, within a window of 365 days before the patient entered the ICU and up to 90 days after the discharge (see Fig. 1). This was completed using a unique patient across the MIMIC-CXR-JPG and MIMIC-IV datasets. This produced a dataset of 2,774 multimodal samples, each sample contains chest radiographs and ICU vital sign values, laboratory results and patient metadata. For more details see [10].

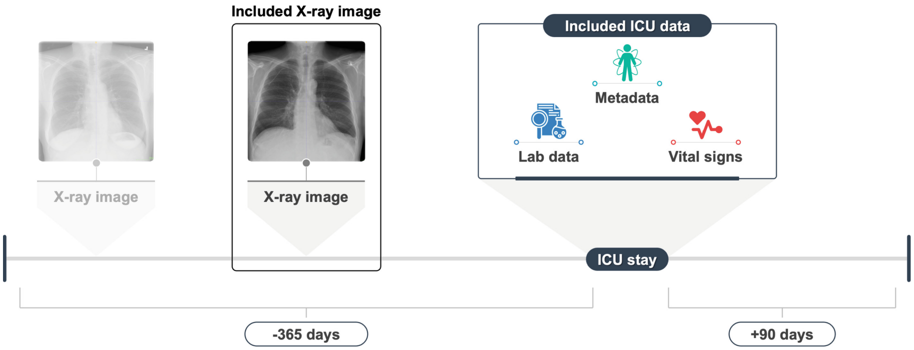


Fig. 1. We merged the closest radiographic study within a time window of 365 days prior to ICU admission and 90 days after release to the data collected during the patient's ICU stay.

2.3 Pre-processing

Imaging Data. To improve model robustness images were re-sized to squared images (244 pixels for cardiomegaly prediction, 256 for segmentation) and pixel values were normalized before input to models under both test and train conditions. Under train conditions only, we also performed standard data augmentation steps including random vertical and horizontal flips and random rotation up to 10°.

Non-imaging Data. Patient metadata was combined with time-series data such as vital sign recordings, using summary statistics such as minimum, maximum, and mean values.

2.4 Models

Heart and Lung Detection. For the detection of hearts and lungs from chest radiographs, we implemented a Faster R-CNN [26] architecture with a ResNet-50 backbone [12] which was pre-trained on ImageNet [5]. Faster R-CNN has shown previously to perform well in clinical object detection tasks [29]. Independent models were trained for heart and lung detection, each model was trained for 300 epochs using the Adam optimiser [20] with a learning rate reduced by a factor of 0.5 on validation intersection over union (IoU) loss plateau. The model iteration with the lowest validation IoU loss was saved.

Heart and Lung Segmentation. For the segmentation of heart and lungs from chest radiographs we implemented a Mask R-CNN [11] architecture with a ResNet-50 backbone which was pre-trained on ImageNet. Mask R-CNN architectures have shown to provide good results in clinical segmentation tasks [6]. For detection, independent models were trained for heart segmentation and lung segmentation. Each model was trained for 300 epochs using the Adam optimiser and with a learning rate which reduced by a factor of 0.5 on validation IoU loss plateau. In order for the loss to be comparable to the detection, bounding boxes were used to calculate IoU loss. To find the bounding boxes the output masks were made into binary masks using Otsu thresholding [22]. The model iteration with the lowest validation IoU loss was saved. In order to have a metric to evaluate the masks, Dice loss [21] was also calculated for predicted masks.

Cardiomegaly Classification with Non-imaging Data. For cardiomegaly classification using non-imaging ICU data as well as the derived digital biomarkers CTR and CPAR (all stored in tabular format), we implemented XGBoost algorithms [4]. XGBoost is known to perform well for similar classification tasks, especially on sparse data [3,24]. A weighted cross-entropy loss was implemented for training. For these XGBoost models we optimised learning rate, maximum tree depth, and tree sub-sample (the fraction of the database sampled to train each tree) through grid search. The XGBoost model which used only CTR and CPAR values as features had a lower max tree depth, than XGBoost models using the ICU non-imaging data. The exact numerical values of the model hyperparameters can be found in Table 1.

Cardiomegaly Classification with Imaging Data. For cardiomegaly classification using images only, we implemented a ResNet-50 architecture pre-trained on ImageNet. This architecture was shown to provide state-of-the-art results with radiology classification tasks, achieving 0.84 accuracy in classifying cardiomegaly on the CheXpert database [1]. The ResNet-50 algorithm uses a cross-entropy loss function with an Adam optimizer and cyclical learning rates [32]. The network was trained in two stages, for the first 15 epochs we trained only the fully connected layers, before unfreezing the convolutional layers and training the full network at a lower maximum learning rate as the optimal maximum learning rate bounds vary [32]. The numerical values for the learning rate bounds can be found in Table 1.

Cardiomegaly Classification with Multimodal Imaging and Non-imaging Data. For classification of cardiomegaly using the multimodal dataset we implemented the network structure proposed in Grant et al. [10]. This architecture combines imaging (chest radiographs) and non-imaging data (metadata, vital sign values, laboratory values and digital biomarkers) by concatenating outputs of the X-ray feature block and the ICU feature block into the joint feature block (shown in Fig. 2). This method has provided good performance [10] and was used to integrate the digital biomarkers, CTR and CPAR, into the classification process. The training was again completed using the Adam optimizer, cyclical learning rates, binary cross-entropy and was completed in two stages. The convolutional layers of the ResNet in the X-ray feature block are frozen and all fully connected layers in the network are trained for 15 epochs with cyclical learning rates. Once the ResNet layers were then unfrozen, the model was trained for 45 epochs with cyclical learning rates using a lower max learning rate. As above, the numerical values for the learning rate bounds and other parameters describing the multimodal network can be found in Table 1. The concatenation layer used to merge the two modalities uses 32 nodes from the X-ray feature block and 16 nodes from the ICU feature block.

2.5 CTR and CPAR: Computation and Model Integration

To compute the CTR and CPAR values, we trained the segmentation and detection models described in Sect. 2.4. To do this we split the CTR dataset into train (80%), validation (10%) and test (10%) subsets; each subset containing a consistent proportion of the JSRT, Montgomery County, and MIMIC databases. Individual models were trained for heart detection, heart segmentation, lung detection and lung segmentation.

CTR Computation. The CTR was computed as the ratio of the widths of the cardiac and pulmonary bounding boxes. To find cardiac and pulmonary bounding boxes we investigated four methods: detection, segmentation, best score ensemble, and average ensemble. For the detection method, bounding boxes output by the Faster R-CNN models were used directly. For the segmentation method, masks output by the Mask R-CNN models were first passed though Otsu thresholding to give a binary mask. From the binary masks cardiac and pulmonary

Table 1. Numerical values of hyperparameters used in different models.

Model	Hyperparameter	Value
XGBoost	Learning rate	0.1
	Tree sub-sample	0.75
	Max tree depth	8 (Tabular models)
		3 (CTR only model)
ResNet	Network depth	ResNet-50
	Learning rate bounds	1e−05–1e−02 (stage 1)
		2e−05–1e−03 (stage 2)
Multimodal	CNN depth	ResNet-50
	ICU Network size	3 Fully connected layers
	Learning rate bounds	1e−05–1e−02 (stage 1)
		2e−05–1e−03 (stage 2)

bounding boxes for the heart and the lungs were found. For the best score ensemble method each sample was passed through both the Faster R-CNN and Mask R-CNN models. The cardiac and pulmonary predictions with the highest score were then selected as the final heart/lung bounding box. For the average ensemble method, each X-ray image was passed through both the Faster R-CNN and Mask R-CNN models. The cardiac and pulmonary predictions were found by producing a point wise average of the bounding box corner coordinates. The methods with the highest IoU on the test set were used to select the final cardiac and pulmonary bounding boxes in the multimodal network.

CPAR Computation. CPAR was computed using the area of the Otsu thresholded masks produced from the segmentation models. The areas of binary masks were used as the cardiac and pulmonary areas and the CPAR was calculated by finding the ratio of the two areas.

Integration of CTR and CPAR into the Multimodal Network. To combine the image-derived digital biomarkers CTR and CPAR with the cardiomegaly classifiers the two methods described above were integrated into the pre-processing stage of our multimodal approach as shown in Fig. 2. CTR and CPAR are combined with the pre-processed ICU data and passed either to the XGBoost models (non-imaging data only) or to the multimodal network via a feedforward neural network in the ICU feature block.

When training the various modality combinations for cardiomegaly classification we used 5-fold stratified cross-validation, each fold is independent of the others with no image repeated between folds and each fold having a similar positive/negative label distribution. When training each combination, four folds were combined for train data and the last fold was split in half for the validation and test data.

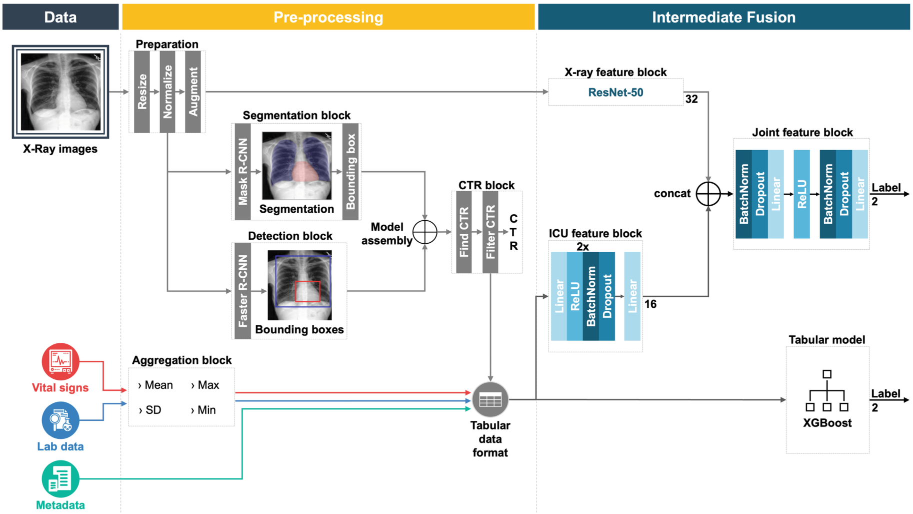


Fig. 2. In the pre-processing step, data is split into tabular (non-imaging ICU data and digital biomarkers) and imaging formats. To generate the biomarkers radiographs are normalised and passed into the segmentation and detection blocks to extract cardiac and pulmonary bounding boxes and masks; then, the CTR bock uses the results from assembled model to find the CTR and CPAR values. The digital biomarker values (CTR and CPAR) are subsequently combined with the non-imaging ICU data (metadata, vital signs, and lab results) into a tabular format. Next, the image data and the combined tabular data are handled either by intermediate or early fusion approach. For the intermediate fusion approach, the image data is augmented and features are extracted by a ResNet-50 in the X-ray feature block. The tabular data is handled via a feedforward neural network in the ICU feature block. Finally, the imaging features are combined with the non-imaging features in the joint feature block via a concatenation layer. Alternatively, for the early fusion approach, predictions are obtained from the tabular data alone via an XGBoost model.

3 Results

3.1 CTR Computation

The performance of the R-CNN model configurations on the CTR test sets is summarised in Table 2 in the form of average IoU scores calculated on the bounding boxes and average precision scores (i.e. area under the curve on a smoothed precision-recall curve) at threshold IoU values of 0.75, 0.85, and 0.95. Additionally, the Dice scores of heart and lung Mask R-CNN models on test sets using the thresholded binary masks were calculated; the heart and lung segmentation models having Dice scores of 0.906 and 0.937, respectively.

For cardiac bounding boxes the best score ensemble model showed the strongest performance with an average IoU score of 0.836 over the test set. For pulmonary bounding boxes the strongest model is the averaged prediction ensemble model with an average IoU score of 0.908 over the test set. As such, these two ensemble models were integrated in the multimodal cardiomegaly classification network to find cardiac and thoracic widths and CTR values. An example of output predictions by each model type and by a combination of the best models is given in Fig. 3.

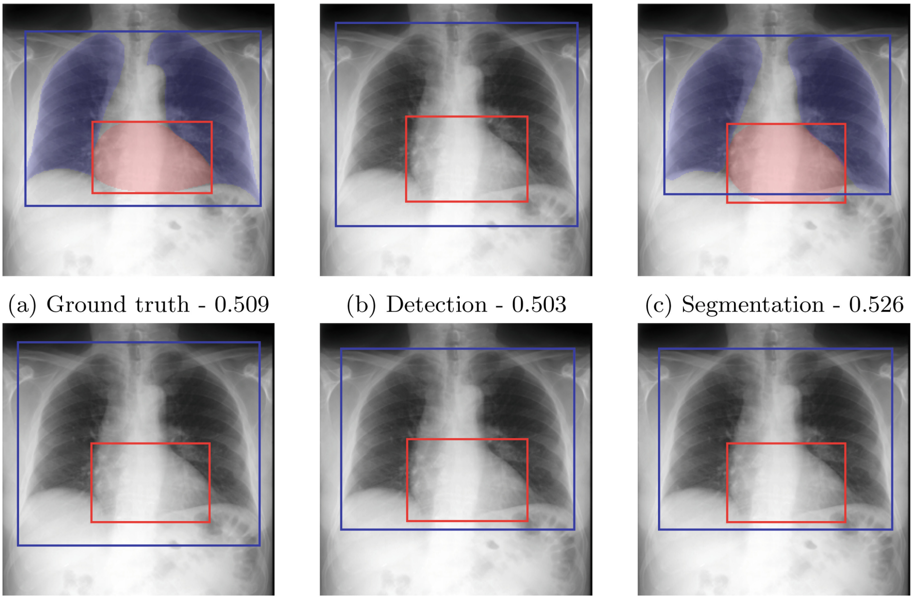
Table 2. IoU score and AP at IoU thresholds of 0.75, 0.85, and 0.95 for bounding boxes found using Fast R-CNN, Mask R-CNN, and ensemble models on test data.

Model	IoU score	AP@0.75	AP@0.85	AP@0.95
Heart Detection	0.810	0.900	0.398	0.020
Heart Segmentation	0.834	0.963	0.678	**0.028**
Heart Ensemble (best)	**0.836**	**0.966**	**0.681**	**0.028**
Heart Ensemble (avg)	0.833	0.954	0.572	–
Lungs Detection	0.853	0.970	0.636	0.100
Lungs Segmentation	0.894	**1.0**	0.963	0.088
Lungs Ensemble (best)	0.852	0.970	0.638	0.100
Lungs Ensemble (avg)	**0.908**	**1.0**	**0.938**	**0.218**

3.2 Multimodal Classification

The performance of models with and without the image-derived digital biomarkers CTR and CPAR are summarised in Table 3 using accuracy (Acc), F1-Score (F1), and area under the receiver operating characteristic curve (AUC). The results scores are averaged over 5-fold cross-validation.

The XGBoost model on non-imaging ICU data showed a distinctly weaker performance (72.4% accuracy) compared to the ResNet-50 on imaging data only (81.9% accuracy), and multimodal network using imaging and non-imaging ICU data (81.9% accuracy). All models which included the digital biomarkers (CTR and CPAR) had comparable performance with accuracy ranging from 81.0% (multimodal network with digital biomarkers) to 82.1% (non-imaging ICU data with digital biomarkers). Overall, all models which included image derived information, either in the form of the digital biomarkers, or direct input of images, had similar level of performance with a accuracy range of 1.1%.



(a) Ground truth - 0.509

(b) Detection - 0.503

(c) Segmentation - 0.526

(d) Best score - 0.491

(e) Averaged - 0.514

(f) Best models - 0.508

Fig. 3. Example output of different methods for finding the bounding boxes and their corresponding calculated CTR values. The best output uses the best score model for cardiac bounding box and averaged model for pulmonary bounding boxes.

Table 3. Performance for different modality combinations with and without digital biomarkers, values averaged over 5-fold cross-validation, with standard deviation in brackets. **includes both digital biomarkers, CTR and CPAR. **includes images, ICU data, and digital biomarkers*

Data used	Model type	Acc	F1	AUC
Images	ResNet-50	0.819 (0.015)	0.874 (0.010)	0.767 (0.029)
ICU data	XGBoost	0.723 (0.030)	0.807 (0.022)	0.651 (0.034)
Images + ICU data	Multimodal	0.819 (0.017)	0.873 (0.013)	0.768 (0.018)
CTR*	XGBoost	0.814 (0.014)	0.859 (0.011)	0.810 (0.012)
ICU data + CTR*	XGBoost	0.821 (0.019)	0.860 (0.012)	0.813 (0.011)
All**	Multimodal	0.810 (0.012)	0.872 (0.009)	0.732 (0.014)

4 Discussion

4.1 Principal Findings

In this work, we considered the classification of cardiomegaly by combining domain knowledge digital biomarkers CTR and CPAR with imaging and non-imaging ICU data. Our results suggest that the multimodal and image-based models are unable to extract additional information beyond what is captured by models trained on CTR and CPAR only. Thus, in the context of cardiomegaly, complicated black-box models may be replaced with carefully curated digital biomarkers, which convey critical clinical information.

4.2 Comparison to Related Works

Sogancioglu et al. [33] compared the predictive power of a black box image classifier to a CTR-based model for cardiomegaly classification, and achieved state of the art results with the later. The CTR-based model outperformed the classic image classification, with AUC values of 0.977 and 0.941, respectively. The conclusions presented in Sogancioglu et al., are in line with the results produced in this work, as the XGBoost model using only CTR and CPAR had an AUC of 0.810, outperforming the ResNet-50 (AUC of 0.767). However, there is a significant difference in the performance of the models in this work compared to their counterparts in Sogancioglu et al., this may be attributed to larger training datasets, cleaner data, and better models. Firstly, the classifiers in Sogancioglu et al. were trained on $65,205$ samples, this contrast to the $2,774$ samples used in this work. Additionally, Sogancioglu et al. excluded any samples where the cardiac boundary was difficult to find, these samples were a notable cause of misclassifications in this work. Lastly, Sogancioglu et al. claimed that the quality of the segmentation models is the most important factor in determining the performance of the CTR-based classifier. Their models performed well achieving IoU scores of 0.87 and 0.95 for heart and lung segmentation respectively. This compares favourably to the IoU scores achieved in this work, 0.836 for heart and 0.907 for lung detection. This may partially be attributed to a larger amount of higher quality data, as Sogancioglu et al. again excluded challenging samples; they used a total of 778 filtered samples to train the segmentation models, compared to 585 unfiltered samples in this work.

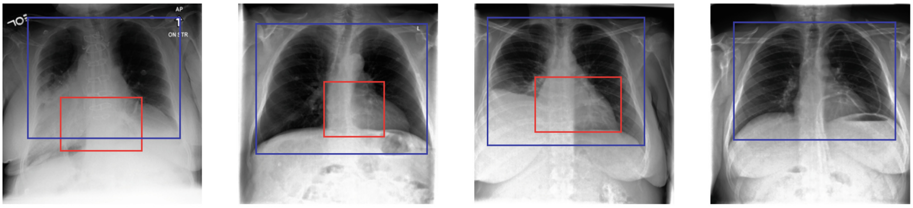
For multimodal classification of cardiomegaly, Grant et al. [10] is most relevant, comparing unimodal and multimodal approaches. The multimodal approaches included images, patient metadata, as well as extensive ICU data. The multimodal approach (accuracy of 0.837 and AUC of 0.880) marginally outperformed the image-only ResNet-50 (accuracy of 0.797 and AUC of 0.840), and greatly outperformed the non-imaging only model (accuracy of 0.700 and AUC of 0.684). Results achieved by Grant et al. are comparable to the results from this work, as the multimodal approaches outperform unimodal ones, with a large drop in performance if no images or image-derived data is included (i.e. no raw images or image-derived biomarkers).

4.3 Strengths and Weaknesses of the Study

Automated Image Classification. An advantage of automatic image labelling tools is that they often avoid cognitive biases. For instance, after making an initial pathological finding on a radiograph, a clinician is less likely to identify further pathological features - a form of premature conclusion bias. Since automatic tools are not subject to this bias, their addition to clinical workflows may increase the pick-up rates of secondary pathologies. In the specific context of this paper, cardiomegaly may be an indicator of many underlying cardiac pathologies and is associated with higher short-term mortality [34], hence the early identification of cardiomegaly is vital. The automatic identification of cardiomegaly can therefore serve as preventative care and a screening tool for cardiac pathologies.

Domain-Based Digital Biomarkers. The CTR (and CPAR) are a clinically valuable diagnostic tools showing high performance when used with classification models (e.g. XGBoost). Since they contain clinically relevant information, their use alongside patient medical data allows the models to more closely imitate the holistic approach taken by clinicians and more accurately reflects existing diagnostic pathways. Hence, leading to a higher degree of confidence in model predictions.

Misclassification Errors. We estimated the digital biomarkers CTR and CPAR values using Mask R-CNN and Fast R-CNN models. Figure 4 shows common cases of false positives and negatives from the XGBoost model trained using the digital biomarkers. The common causes of misclassification are interference from other pathologies leading to inaccurate cardiac and thoracic widths and models' failures to accurately identify the heart.



(a) False negative (b) False negative (c) False positive (d) False positive

Fig. 4. False positive and false negative classifications from XGBoost model trained using only digital biomarkers. Suspected causes of error for the respective images are: (a) area of parenchymal opacity around heart hides heart boundary leading to inaccurate cardiac bounding box; (b) R-CNN models failed to correctly identify image-right boundary of heart; (c) pleural effusion in image-left lung causing pulmonary bounding box width to be smaller than thoracic width; (d) R-CNN models failed to identify heart.

Label Errors. The cardiomegaly labels were derived from the free text reports associated with chest radiographs. These labels may contain errors since the radiographs alone may be insufficient for definitive cardiomegaly diagnosis. Also, the automatic label extraction from the free test reports may be another source of error [13,23]. It is known that these procedures can introduce noise and affect model performance [18]. We took steps to mitigate these errors by employing the procedures described in Sect. 2.2.

Acknowledgements. BWP acknowledges Rutherford Fund at Health Data Research UK (HDR UK) (MR/S004092/1) and Nuffield Department of Population Health (NDPH) Senior Research Fellowship.

References

1. Bressem, K.K., Adams, L.C., Erxleben, C., Hamm, B., Niehues, S.M., Vahldiek, J.L.: Comparing different deep learning architectures for classification of chest radiographs. Sci. Rep. **10**(1), 13590 (2020)
2. Candemir, S., et al.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans. Med. Imaging **33**(2), 577–590 (2014)
3. Chang, W., et al.: A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics **9**(4), 178 (2019)
4. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. KDD 2016. ACM, New York, NY, USA (2016)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Kai, L., Li, F-F.: ImageNet: a large-scale hierarchical image database. In: Institute of Electrical and Electronics Engineers (IEEE), pp. 248–255 (2010)
6. Durkee, M., Abraham, R., Ai, J., Fuhrman, J., Clark, M., Giger, M.: Comparing mask r-CNN and u-net architectures for robust automatic segmentation of immune cells in immunofluorescence images of lupus nephritis biopsies. In: Leary, J., Tarnok, A., Georgakoudi, I. (eds.) Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XIX. SPIE, March 2021
7. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX **7**, 100864 (2020)
8. van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med. Image Anal. **10**(1), 19–40 (2006)
9. Goldberger, A.L., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
10. Grant, D., Papież, B., Parsons, G., Tarassenko, L., Mahdi, A.: Deep learning classification of cardiomegaly using combined imaging and non-imaging ICU data. In: Medical Image Understanding and Analysis, pp. 547–558. Springer International Publishing, July 2021. https://doi.org/10.1007/978-3-030-80432-9_40
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem, pp. 770–778. IEEE Computer Society (2016)
13. Irvin, J., et al.: CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: 33rd AAAI Conference on Artificial Intelligence. AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, vol. 33, pp. 590–597. AAAI Press (2019)

14. Ishida, T., Katsuragawa, S., Chida, K., MacMahon, H., Doi, K.: Computer-aided diagnosis for detection of cardiomegaly in digital chest radiographs. In: Medical Imaging 2005: Image Processing, vol. 5747, p. 914. SPIE (2005)
15. Jaeger, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE Trans. Med. Imaging **33**(2), 233–245 (2014)
16. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV v0.4. Tech. rep., MIT Laboratory for Computational Physiology (2020)
17. Johnson, A.E.W., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 1–8 (2019)
18. Johnson, A.E.W., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv (2019)
19. Karaman, S.: Cardiothoracic area ratio for evaluation of ejection fraction in patients. J. Clin. Anal. Med. **10**, 188–192 (2019)
20. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR (2015)
21. Milletari, F., Navab, N., Ahmadi, S.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
22. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
23. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: NegBio: a high-performance tool for negation and uncertainty detection in radiology reports (2017)
24. Pimentel, M.A.F., et al.: Detecting deteriorating patients in hospital: development and validation of a novel scoring system. Am. J. Respir. Crit. Care Med. **204**, 44–52 (2021)
25. Que, Q., et al.: CardioXNet: automated detection for cardiomegaly based on deep learning. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2018-July, pp. 612–615. Institute of Electrical and Electronics Engineers Inc. (2018)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks (2015)
27. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ **359**, j4683 (2017)
28. Rosman, D., et al.: Imaging in the land of 1000 hills: Rwanda radiology country report. J. Glob. Radiol. **1**(1), 5 (2015)
29. Sa, R., et al.: Intervertebral disc detection in x-ray images using faster R-CNN. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, July 2017
30. Samak, Z.A., Clatworthy, P., Mirmehdi, M.: Prediction of thrombectomy functional outcomes using multimodal data. In: Papież, B.W., Namburete, A.I.L., Yaqub, M., Noble, J.A. (eds.) MIUA 2020. CCIS, vol. 1248, pp. 267–279. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52791-4_21
31. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule. Am. J. Roentgenol. **174**(1), 71–74 (2000)
32. Smith, L.N.: Cyclical learning rates for training neural networks. In: Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, pp. 464–472. Institute of Electrical and Electronics Engineers Inc. (2017)
33. Sogancioglu, E., Murphy, K., Calli, E., Scholten, E.T., Schalekamp, S., Van Ginneken, B.: Cardiomegaly detection on chest radiographs: segmentation versus classification. IEEE Access **8**, 94631–94642 (2020)

34. Yen, T., Lin, J.L., Lin-Tan, D.T., Hsu, K.H.: Cardiothoracic ratio, inflammation, malnutrition, and mortality in diabetes patients on maintenance hemodialysis. Am. J. Med. Sci. **337**(6), 421–428 (2009)