SEGMENTATION-DERIVED CARDIOTHORACIC FEATURES FOR A PILOT CARDIOMEGALY SCREENING WORKBENCH ON CHEST RADIOGRAPHS

Author: Rungphob Lertvilaivithaya

Year: 2026

Project Title: Cardiomegaly AI X-Ray Screening Workbench

Research area: medical imaging, machine learning, and applied computer science

Note: This paper describes a research prototype only. It is not a clinical diagnostic device.

TABLE OF CONTENTS

ABSTRACT

Cardiomegaly refers to enlargement of the cardiac silhouette on chest radiography and is commonly evaluated using the cardiothoracic ratio (CTR), defined as the ratio of maximal cardiac width to maximal internal thoracic width. On an appropriately acquired posteroanterior radiograph, a CTR greater than 0.50 may indicate cardiomegaly; however, its interpretation is influenced by projection type, patient rotation, inspiratory effort, positioning, and incomplete pulmonary expansion [1–3]. These technical and anatomical factors may produce apparent cardiac enlargement, thereby limiting the reliability of automated image-based screening.

Machine-learning models trained on public chest-radiograph datasets may also be affected by uncertainty associated with report-derived labels. In small-sample studies, label noise, sampling bias, and heterogeneity in image acquisition may disproportionately influence estimated model performance and generalisability [4,5].

This study proposes an interpretable screening workflow that identifies relevant anatomical structures and derives cardiothoracic measurements rather than relying exclusively on raw-image classification. Because radiographic assessment of cardiomegaly is fundamentally based on anatomical proportions, this feature-based approach was designed to support transparent preliminary screening in settings with limited positive training data [1,2].

Keywords: cardiomegaly; chest radiograph; cardiothoracic ratio; segmentation; calibrated support vector classifier

CHAPTER 1

Introduction

1.1 Research Background

Cardiomegaly refers to enlargement of the cardiac silhouette observed on chest radiography. On frontal chest radiographs, it is commonly assessed using the cardiothoracic ratio (CTR), which compares the maximum transverse cardiac diameter with the maximum internal transverse diameter of the thoracic cavity. On a properly acquired posteroanterior (PA) chest radiograph, a CTR above approximately 0.50 may indicate an enlarged cardiac silhouette, although this threshold should be interpreted within the relevant clinical and imaging context [1,2].

Although the CTR is conceptually straightforward, its interpretation is affected by technical and anatomical factors. Measurement is most reliable on upright PA radiographs because anteroposterior (AP) projections can magnify the cardiac silhouette and make the heart appear larger than its true size [3]. Patient rotation, reduced inspiratory effort, positioning, and incomplete lung expansion may also alter the apparent relationship between the cardiac silhouette and thoracic cavity [1,2]. These sources of variation create an important challenge for automated screening systems because a model must distinguish patterns consistent with cardiomegaly from apparent enlargement caused by image acquisition or positioning.

Public chest-radiograph datasets have supported the development of machine-learning methods in thoracic imaging. However, several commonly used datasets rely partly or primarily on labels extracted from radiology reports using natural-language-processing methods rather than direct image-level adjudication for every image. For example, ChestX-ray8 used text-mined disease labels from associated radiology reports, whereas CheXpert used an automated labeler to identify observations and uncertainty from reports [4,5]. These labelling methods can introduce uncertainty because a report-derived label may not perfectly correspond to the radiographic appearance in an individual image. This limitation is especially relevant in small-sample studies, where label noise, sampling bias, and differences in image acquisition may substantially influence reported performance [4,5].

This study investigates an anatomy-aware approach to cardiomegaly screening using segmentation-derived measurements of the heart and lungs. Rather than relying solely on raw-image classification, the proposed workflow first estimates relevant anatomical structures and subsequently derives interpretable cardiothoracic features. This design was selected because radiographic assessment of cardiomegaly is closely related to measurable anatomical proportions, particularly the relationship between the cardiac

silhouette and thoracic width [1,2]. A feature-based approach may therefore be more appropriate than high-capacity image classification when the available positive cohort is limited and interpretability is required.

1.2 Research Question and Objectives

The central research question of this study was whether segmentation derived cardiothoracic features can support pilot level cardiomegaly screening on frontal chest radiographs using patient level internal testing.

The study aimed to develop a machine learning workflow that extracts interpretable anatomical features from frontal chest radiographs and uses these measurements to train a screening oriented classifier. A further objective was to reduce the risk of data leakage by excluding bounding box coordinates from model inputs and by separating the dataset at the patient level. The study also evaluated internal model performance using standard classification metrics, including sensitivity, specificity, precision, accuracy, F1 score, ROC AUC, and the confusion matrix.

In addition to model development, the study examined how the trained classifier could be integrated into a structured research workflow. The prototype system was designed to report probability, threshold information, selected anatomical feature values, and review oriented output. The objective was not to create an autonomous diagnostic device, but to demonstrate how an interpretable model could be connected to a transparent screening workflow. The study also sought to define the limitations of the dataset, reference standard, threshold selection process, and internal testing design.

1.3 Significance of the Study

The significance of this study lies in the development of an interpretable pilot workflow for cardiomegaly screening rather than the establishment of a clinically validated diagnostic system. Cardiomegaly assessment on chest radiography has an identifiable anatomical basis, as the cardiothoracic ratio is derived from the relationship between maximal cardiac width and internal thoracic width on a posteroanterior radiograph [1,2]. This relationship makes cardiomegaly an appropriate application for investigating whether segmentation-derived anatomical measurements and feature-based machine-learning methods can support preliminary image-based screening.

Unlike an exclusively image-based classification model, the proposed workflow links its predictions to measurable radiographic features, including cardiac and thoracic dimensions. This design improves transparency by allowing users to examine the anatomical measurements underlying the model output. Such transparency is particularly important in medical-imaging artificial intelligence, where reporting guidance

emphasises clear documentation of datasets, model development, evaluation procedures, intended use, and study limitations [11,12].

The findings should be interpreted conservatively. The dataset was limited in size, labels were operational rather than independently clinician-adjudicated, and the classification threshold was refined during internal model development. Consequently, the reported performance represents internal feasibility evidence and should not be interpreted as evidence of diagnostic accuracy, clinical utility, or readiness for real-world implementation. Current guidance for clinical prediction models and medical-imaging AI likewise emphasises transparent reporting and evaluation using independent external datasets before claims of generalisability can be made [11,12].

Accordingly, the principal contribution of this study is methodological and practical. Methodologically, it evaluates whether interpretable cardiothoracic measurements can provide a meaningful internal screening signal in a limited-data setting. Practically, it demonstrates how these measurements may be incorporated into a structured research workbench that reports prediction probabilities, decision thresholds, anatomical feature values, and review-oriented outputs. Further investigation using larger, independently collected, and clinician-adjudicated datasets is necessary to determine the reproducibility, generalisability, and potential clinical relevance of the proposed approach.

CHAPTER 2

Literature Review

Cardiomegaly assessment on chest radiography has traditionally relied on the visual and geometric relationship between the cardiac silhouette and the thoracic cavity. One of the most widely used radiographic measurements is the cardiothoracic ratio (CTR), defined as the ratio of the maximal transverse cardiac diameter to the maximal internal transverse thoracic diameter. An increased CTR may suggest enlargement of the cardiac silhouette; however, its interpretation depends substantially on radiographic projection and image-acquisition conditions [1,2]. Cardiothoracic measurements are generally most reliable on standard upright posteroanterior radiographs, whereas portable anteroposterior radiographs may exaggerate the apparent size of the heart because of magnification and differences in patient positioning [1–3].

The interpretation of cardiomegaly is therefore not based on a single numerical measurement alone. Radiologists also consider lung volume, patient rotation, projection, exposure, positioning, and the overall technical adequacy of the radiograph [1–3]. Reduced inspiratory effort can decrease visible lung expansion and make the cardiac silhouette appear proportionally larger. Similarly, patient rotation or incomplete positioning may alter mediastinal alignment and the measured thoracic width. These limitations are particularly relevant to automated systems because a model that responds only to apparent cardiac width may misclassify technically suboptimal images as cardiomegaly.

This concern is supported by McKee and Ferrier, who compared cardiomegaly reported on chest radiographs with echocardiographic evidence of cardiac enlargement [6]. In their study of 244 patients, chest-radiograph assessment demonstrated a sensitivity of 40.0%, specificity of 91.0%, positive predictive value of 56.0%, and negative predictive value of 84.0% when echocardiography was used as the reference standard [6]. These findings indicate that an enlarged cardiac silhouette on chest radiography does not consistently correspond to echocardiographically defined cardiomegaly. Accordingly, chest-radiograph findings should be interpreted as a screening indication for further cardiac assessment rather than definitive evidence of anatomical cardiac enlargement [6].

Public chest-radiograph datasets have played an important role in the development of machine-learning methods for thoracic imaging. Large retrospective image collections have enabled researchers to train and evaluate automated systems for detecting radiographic findings at a scale that would otherwise be difficult to achieve. However, many public datasets use labels derived from radiology reports, automated natural-language-processing methods, or selected annotation files rather than

independent clinician adjudication for every image [4,5]. Such labels remain valuable for research but should not be interpreted as infallible diagnostic ground truth. Label uncertainty is particularly important in small pilot studies, in which a limited number of mislabeled cases may substantially affect estimates of sensitivity, specificity, predictive value, and overall model performance.

Machine-learning approaches to chest-radiograph analysis range from end-to-end deep-learning systems to feature-based classifiers. Deep neural networks can learn complex visual representations directly from image pixels; however, they generally require large and heterogeneous datasets to reduce the risk of overfitting, shortcut learning, and poor generalisability. In contrast, feature-based approaches use predefined measurements that represent clinically meaningful image characteristics. For cardiomegaly screening, this approach is appropriate because the target finding is closely associated with measurable cardiac and thoracic geometry [1,2,6]. Segmentation-derived variables, including estimated heart width, thoracic width, CTR, heart area, lung area, and heart-to-lung area relationships, may therefore provide a more interpretable basis for classification than a model operating exclusively on raw image pixels.

Recent research has shown that artificial intelligence may support cardiomegaly assessment by automatically deriving CTR-related measurements from chest radiographs. Ajmera et al. developed a deep-learning model for automated CTR calculation and evaluated radiologist performance before and after access to the AI-generated measurement [7]. The radiologist's sensitivity for identifying cardiomegaly increased from 40.5% without AI assistance to 88.4% with AI-generated CTR information, although the improvement was accompanied by an increase in false-positive classifications [7]. This result suggests that AI-derived anatomical measurements may be useful as a second-reader or decision-support tool, particularly when visual interpretation alone may overlook borderline cases.

Similarly, Saiviroonporn et al. evaluated an AI-assisted CTR measurement workflow using a large clinical dataset of chest radiographs [8]. Their system reduced the time required for CTR measurement while producing outputs that were accepted without additional adjustment in most evaluated images [8]. These findings support the potential value of segmentation-based measurements for improving efficiency and consistency in radiographic assessment. However, such systems are intended to support clinician interpretation rather than replace clinical judgement or establish a definitive diagnosis independently.

Classical machine-learning algorithms, including logistic regression, support vector classifiers, and tree-based ensemble models, are suitable for structured feature tables

containing anatomical measurements. These methods may be particularly useful when the number of available observations is limited and the extracted features already represent clinically meaningful aspects of the radiograph. A calibrated linear support vector classifier is appropriate for an interpretable pilot workflow because it can learn from engineered anatomical features while producing calibrated probability estimates. This allows the system output to be communicated as a screening probability rather than solely as an uncalibrated classification score.

Transparent reporting is essential in medical-imaging artificial-intelligence research. Current reporting guidance emphasises the need to describe the data source, reference standard, preprocessing procedures, model inputs, data partitioning, internal evaluation, external validation, intended use, and study limitations clearly [9,10]. This distinction is particularly important because strong performance on an internal split from one dataset does not demonstrate that a model will generalise to different hospitals, scanner types, patient populations, or image-acquisition protocols. Claims of clinical readiness should therefore be avoided unless a model has been evaluated using independent, representative, and clinician-adjudicated external data [9,10].

The present study adopts this conservative framework. It uses a public retrospective dataset to investigate whether segmentation-derived cardiothoracic features can support pilot-level cardiomegaly screening. The dataset labels are not treated as definitive clinical diagnoses, and the study does not claim external validation or clinical deployment. Instead, the study evaluates whether an anatomy-aware, feature-based model can produce an interpretable internal screening signal and whether the resulting model can be integrated into a structured research workflow. The work should therefore be understood as an early-stage medical-artificial-intelligence feasibility investigation rather than a clinically validated diagnostic system.

Conventional Cardiothoracic Ratio as a Clinical Baseline

A conventional baseline for cardiomegaly screening is radiographic assessment of cardiac size using the cardiothoracic ratio (CTR). The CTR is defined as the ratio of the maximal transverse cardiac diameter to the maximal internal thoracic diameter on a posteroanterior chest radiograph. A CTR greater than 0.50 is conventionally considered suggestive of cardiomegaly; however, it represents an indirect radiographic indicator of cardiac enlargement rather than a direct measurement of cardiac chamber size, ventricular mass, or cardiac volume [7].

Interpretation of the CTR is affected by technical and physiological factors. Anteroposterior projection, suboptimal inspiratory effort, patient rotation, supine positioning, and variation in body habitus may alter the apparent dimensions of the

cardiac silhouette and result in an overestimation or underestimation of cardiac size [7]. Therefore, although CTR measurement is widely recognised and clinically accessible, it should be interpreted as a screening measure rather than a definitive diagnostic assessment of structural cardiac enlargement.

The diagnostic limitations of conventional chest-radiograph assessment were demonstrated by McKee and Ferrier, who compared radiographic cardiomegaly with echocardiographic evidence of cardiac enlargement in 244 patients [8]. Of the 39 patients reported as having cardiomegaly on chest radiography, 22 also demonstrated cardiomegaly on echocardiography. In contrast, 33 of the 55 patients identified as having cardiomegaly on echocardiography were not reported as having cardiomegaly on chest radiography. Using echocardiography as the reference standard, chest radiography demonstrated a sensitivity of 40.0%, specificity of 91.0%, positive predictive value of 56.0%, and negative predictive value of 84.0% [8].

These results indicate that conventional radiographic assessment may be relatively specific but insufficiently sensitive for excluding echocardiographically defined cardiomegaly. Reconstructing the reported diagnostic outcomes yields 22 true-positive cases, 17 false-positive cases, 33 false-negative cases, and 172 true-negative cases. This pattern is clinically relevant because it suggests that visual interpretation of cardiac enlargement on chest radiographs may fail to identify a substantial proportion of patients with structural cardiac enlargement confirmed by echocardiography.

Recent studies have investigated whether artificial intelligence can improve cardiothoracic-ratio assessment and support radiologist interpretation. Ajmera et al. developed a deep-learning model based on an Attention U-Net architecture to automatically calculate the CTR from 1,012 posteroanterior chest radiographs [9]. In an observer-performance experiment, a radiologist first interpreted chest radiographs without artificial-intelligence assistance and subsequently reviewed the images with access to the AI-generated CTR. The radiologist’s sensitivity for cardiomegaly increased from 40.5% without AI assistance to 88.4% with access to the AI-derived CTR, although the number of false-positive classifications also increased [9]. This finding suggests that AI-generated anatomical measurements may be valuable as a second-reader or decision-support tool, particularly for borderline cases that may otherwise be overlooked during visual interpretation.

Further evidence was provided by Saiviroonporn et al., who evaluated an AI-assisted CTR measurement system using a clinical dataset of 9,386 chest radiographs [10]. Their combined deep-learning model produced measurements that users accepted without adjustment in 77.8% of images and reduced mean CTR measurement time from 10.6 seconds to 1.07 seconds per image compared with manual measurement [10]. The

study supports the practical use of AI-assisted CTR measurement to reduce radiologist workload while retaining clinician oversight.

These studies provide a useful benchmark for the present research because they demonstrate that segmentation-derived cardiothoracic measurements can support cardiomegaly screening and radiologist workflow. However, they should not be interpreted as directly comparable with the present pilot results. McKee and Ferrier used echocardiography as the reference standard, whereas the present study used operational labels from a public chest-radiograph dataset. Similarly, the AI studies were evaluated using substantially larger clinical datasets and radiologist-derived measurements. Future external validation of the proposed workflow should therefore compare the model directly with conventional CTR assessment and radiologist interpretation on the same independently collected dataset, ideally using echocardiography or cardiac magnetic resonance imaging as the reference standard.

CHAPTER 3

Research Method

3.1 Study Design

This study was designed as a retrospective pilot model development study using publicly available frontal chest radiographs. The aim was to evaluate whether segmentation derived anatomical features could support cardiomegaly screening under internal testing conditions. The study was not designed to establish clinical validity or diagnostic readiness. Instead, it examined the feasibility of combining automated anatomical segmentation, interpretable feature extraction, and a screening oriented classifier within a structured research workflow.

The intended use of the model was screening support. In this context, the model output was treated as a signal for human review rather than as an autonomous diagnosis. This distinction was important because the dataset was limited in size, the labels were operational research labels, and the model was evaluated only within an internal patient level split.

3.2 Data Source and Reference Standard

The study cohort was derived from the NIH ChestX-ray dataset. The cardiomegaly positive group consisted of images identified through the dataset's cardiomegaly annotation records, while the comparison group consisted of images labeled as "No Finding" in the available metadata. The active cohort contained 146 cardiomegaly labeled images and 438 metadata defined comparison images, giving a total of 584 images.

The reference standard should be interpreted cautiously. A cardiomegaly positive label indicated that the image was included in the relevant annotation subset, while a comparison image indicated that the dataset metadata did not list a finding. These labels are suitable for pilot model development, but they are not equivalent to independent clinician adjudication. In particular, the "No Finding" label should not be interpreted as proof that the radiograph was normal after expert review.

The dataset was divided into training, validation, and internal test partitions at the patient level. This approach reduced the risk that images from the same patient would appear across multiple splits. The training set contained 408 images, the validation set contained 88 images, and the internal test set contained 88 images. The cardiomegaly and comparison classes were distributed proportionally across these partitions.

Additional checks found no patient overlap, image overlap, or file hash overlap across the training, validation, and internal test sets.

3.3 Image Processing and Feature Extraction

Each image was processed to generate anatomical measurements related to the heart and lungs. A pretrained chest radiograph segmentation model was used to estimate the cardiac and pulmonary regions. These segmentation outputs were then converted into structured features that described the relative size, position, and proportional relationship of the heart and lung fields.

The extracted features included measurements such as segmentation derived cardiothoracic ratio, heart width relative to image width, lung width relative to image width, heart area relative to lung area, and lung symmetry. Additional features described image quality and acquisition related factors, including projection, rotation, inspiration, and exposure proxies. These measurements were selected because apparent cardiac enlargement can be affected not only by true heart size, but also by technical factors that alter the appearance of the thorax.

The model was designed to avoid direct use of bounding box coordinates as predictive inputs. Although annotation records were used to identify the positive pilot cohort, the classifier itself was trained using image derived anatomical measurements, segmentation features, metadata features, and engineered interactions. This design was intended to make the model more consistent with a deployable workflow, where new uploaded images would not arrive with disease bounding box annotations.

3.4 Feature Groups

The final feature set consisted of several broad categories. Thoracic and cardiac width features estimated the relative size of the cardiac silhouette within the chest. Regional opacity features summarized intensity and structural information in the lower central thorax. Projection and quality features represented acquisition conditions that could affect apparent heart size. Segmentation anatomy features measured relationships between the estimated heart and lung masks. Metadata features provided limited contextual information when available. Engineered interaction features combined clinically relevant measurements, such as cardiothoracic proportions with projection or quality indicators.

This feature based design was chosen because cardiomegaly is closely related to measurable anatomical proportions. A feature based model also allows the prediction to be interpreted through recognizable measurements rather than only through raw pixel

patterns. However, this approach depends on the quality of the segmentation masks. If heart or lung segmentation fails, the derived measurements may become unreliable. For this reason, segmentation quality should be considered an important limitation of the pipeline.

3.5 Model Development

Several candidate classifiers were considered during model development, including logistic regression, random forest, Extra Trees, calibrated linear support vector classification, and gradient boosted tree models. The selected active model was a calibrated linear support vector classifier. This model type was appropriate for the study because the dataset was small and the input features were already structured around clinically meaningful measurements.

The active pipeline used preprocessing steps to handle missing values and standardize numeric inputs before classification. Calibration was applied to convert the support vector classifier output into probability like scores. This was necessary because the prototype workflow reported model probability and threshold based interpretation, which would be less meaningful if based only on an uncalibrated decision score.

3.6 Threshold Selection

The model used a sensitivity oriented thresholding strategy because the intended task was screening support. In screening, a false negative represents a missed case that would not be flagged for review, while a false positive represents an additional case sent for human assessment. Therefore, the operating threshold was selected with emphasis on reducing missed cardiomegaly labeled cases while maintaining a manageable number of false positives.

During internal development, a validation based threshold was first identified. When applied to the internal test split, this threshold produced missed positive cases. The threshold was subsequently lowered, which removed the observed false negatives in the internal test split but increased the number of false positives. This adjustment is a major methodological limitation. Because the final threshold was selected after observing internal test behavior, the reported test results should be interpreted as internal development evidence rather than as a fully independent final test estimate.

3.7 Internal Evaluation and Statistical Analysis

Model performance was evaluated on the internal patient level test split using standard classification metrics. These included accuracy, precision, sensitivity, specificity, F1

score, ROC AUC, Brier score, and the confusion matrix. Bootstrap confidence intervals were calculated for selected metrics to provide an estimate of uncertainty.

The internal test set contained only 88 images, including 22 cardiomegaly positive cases. As a result, performance estimates are sensitive to small changes in classification outcome. For example, a single additional false negative would substantially reduce the sensitivity estimate. The statistical results should therefore be treated as preliminary and should not be generalized beyond the internal dataset.

3.8 Leakage Control

Data leakage was addressed through several safeguards. The dataset was split at the patient level to prevent images from the same patient appearing in more than one partition. Image level and file hash checks were also performed to assess duplicate overlap across the training, validation, and internal test sets. These checks found no overlap across the active partitions.

A separate leakage concern involved the use of bounding box annotations. Direct bounding box coordinates were excluded from the active model inputs because such coordinates would provide disease localization information that would not be available during real world inference. The positive cohort was identified using annotation records, but the classifier did not use bounding box coordinates or bounding box prior features as input variables. This distinction is important because it separates cohort definition from deployable model prediction.

These safeguards improve the internal validity of the study, but they do not eliminate all sources of bias. They do not prove that the labels are clinically perfect, that comparison images are truly normal, or that the model will generalize to images from different institutions. They should therefore be understood as leakage controls for internal model development rather than evidence of external validation.

3.9 Prototype Workflow Integration

The trained model was integrated into a prototype screening workflow. The system accepted an uploaded chest radiograph, extracted the required features, generated a cardiomegaly probability, applied the selected threshold, and returned a structured screening result. The output included probability information, threshold based interpretation, review priority language, and selected anatomical feature values.

The prototype also supported structured report generation. This component was included to demonstrate how model output could be communicated in a consistent format rather than as an isolated classification label. The report framing emphasized

screening support and human review, not autonomous diagnosis. In the context of this study, the application should be understood as a translational research artifact that demonstrates workflow feasibility rather than as a clinically validated software device.

3.10 Reproducibility

The study maintained reproducibility through saved model artifacts, split records, feature tables, prediction files, metric summaries, and leakage audit outputs. These files allowed the cohort definition, data partitions, model inputs, predictions, and reported metrics to be reviewed. This is important for transparency because it links the written results to concrete development artifacts rather than relying only on narrative description.

Nevertheless, reproducibility within one project environment is not equivalent to external validity. The current artifacts support internal verification of the pilot workflow, but further evaluation would require a locked model and threshold applied to a new independently labeled cohort. Such testing would be necessary before stronger claims about generalizability or clinical performance could be made.

CHAPTER 4

Results and Discussion

4.1 Main Test Results

The active pilot model was evaluated on an internal patient level test set of 88 radiographs. At the selected operating threshold, the model achieved 92.05% accuracy, 75.86% precision, 100.00% sensitivity, 89.39% specificity, an F1 score of 86.27%, and an ROC AUC of 97.18%. The corresponding confusion matrix showed 59 true negatives, 7 false positives, 0 false negatives, and 22 true positives.

The model's performance reflects a sensitivity oriented screening strategy. At this threshold, all cardiomegaly labeled cases in the internal test split were identified, while 7 comparison images were incorrectly classified as screening positive. This tradeoff is consistent with a screening objective, where missed positive cases are usually treated as more concerning than additional cases referred for human review.

Metric	Value
Accuracy	0.920
Precision / PPV	0.759
Recall / sensitivity	1.000
Specificity	0.894
F1 score	0.863
ROC AUC	0.972
Brier score	0.062

Table 1. Held-out pilot test performance.

Metric	Bootstrap 95% CI
Accuracy	0.864 to 0.977
Precision / PPV	0.600 to 0.906
Recall / sensitivity	1.000 to 1.000
Specificity	0.818 to 0.966
ROC AUC	0.937 to 0.995

Table 2. Bootstrap confidence intervals for selected internal test metrics.

The bootstrap confidence intervals indicate substantial uncertainty, particularly because the internal test set was small. The positive test sample contained only 22 cardiomegaly labeled images, so even one additional false negative would noticeably reduce the sensitivity estimate. For this reason, the results should be interpreted as internal pilot evidence rather than as a stable estimate of real world clinical performance.

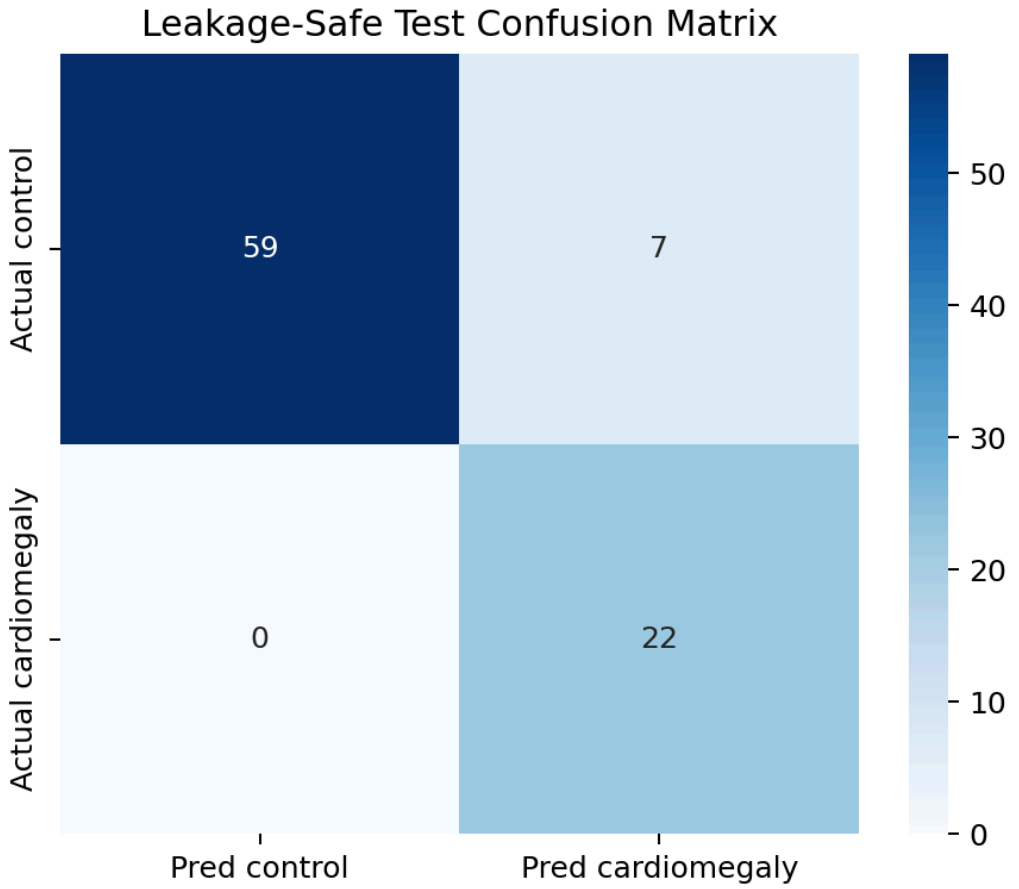


Figure 1. Confusion matrix for the held-out pilot test set.

The confusion matrix provides the clearest summary of the model's behavior at the selected threshold. The absence of false negatives in this split is notable, but it should not be interpreted as evidence that the model would avoid false negatives in future clinical use. A larger independent test cohort is needed to determine whether this sensitivity is reproducible.

4.2 Threshold Tradeoff

The selected threshold was intended to prioritize sensitivity while limiting the number of false positives. Raising the threshold would reduce the number of comparison images classified as positive, but it could also increase the risk of missed cardiomegaly labeled cases. Lowering the threshold would make the model more sensitive, but at the cost of referring more images for review.

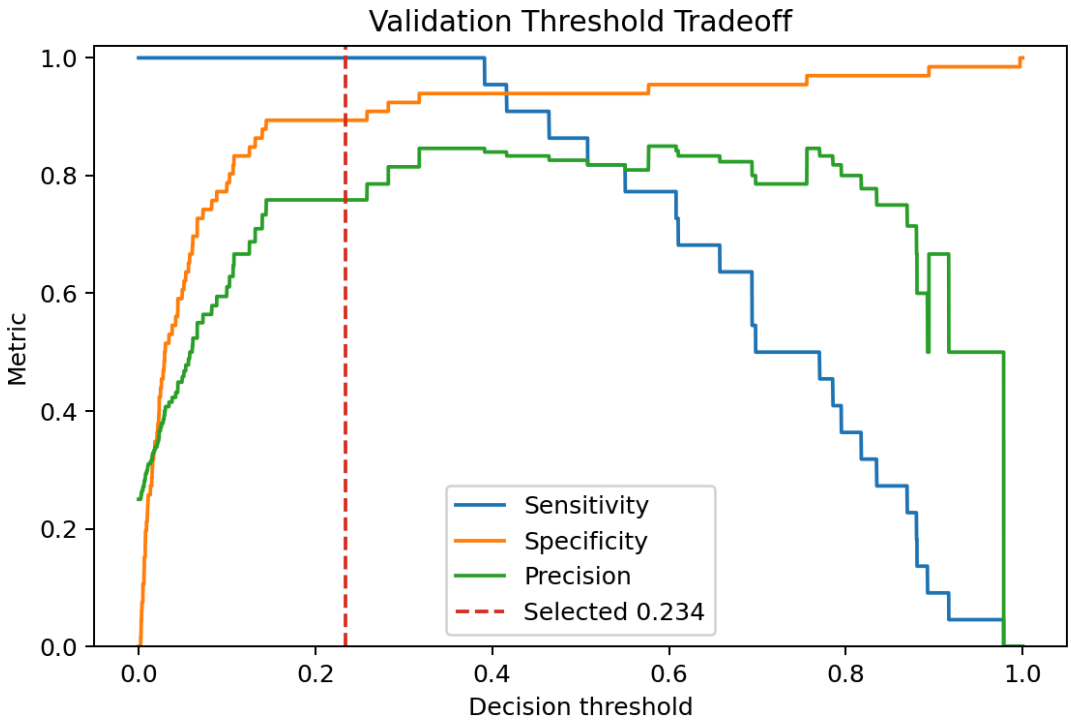


Figure 2. Sensitivity, specificity, and precision across decision thresholds.

The threshold analysis shows this tradeoff directly. The final threshold should be understood as an internal development operating point rather than as a clinically established cutoff. This distinction is important because the threshold was adjusted during internal development after observing test split behavior. As a result, the final internal test performance should not be described as a fully independent validation result.

4.3 ROC Curve and Discrimination

The ROC curve evaluates how well the model ranks cardiomegaly labeled images above comparison images across possible thresholds. The internal test ROC AUC of 97.18% suggests strong apparent discrimination within the study cohort.

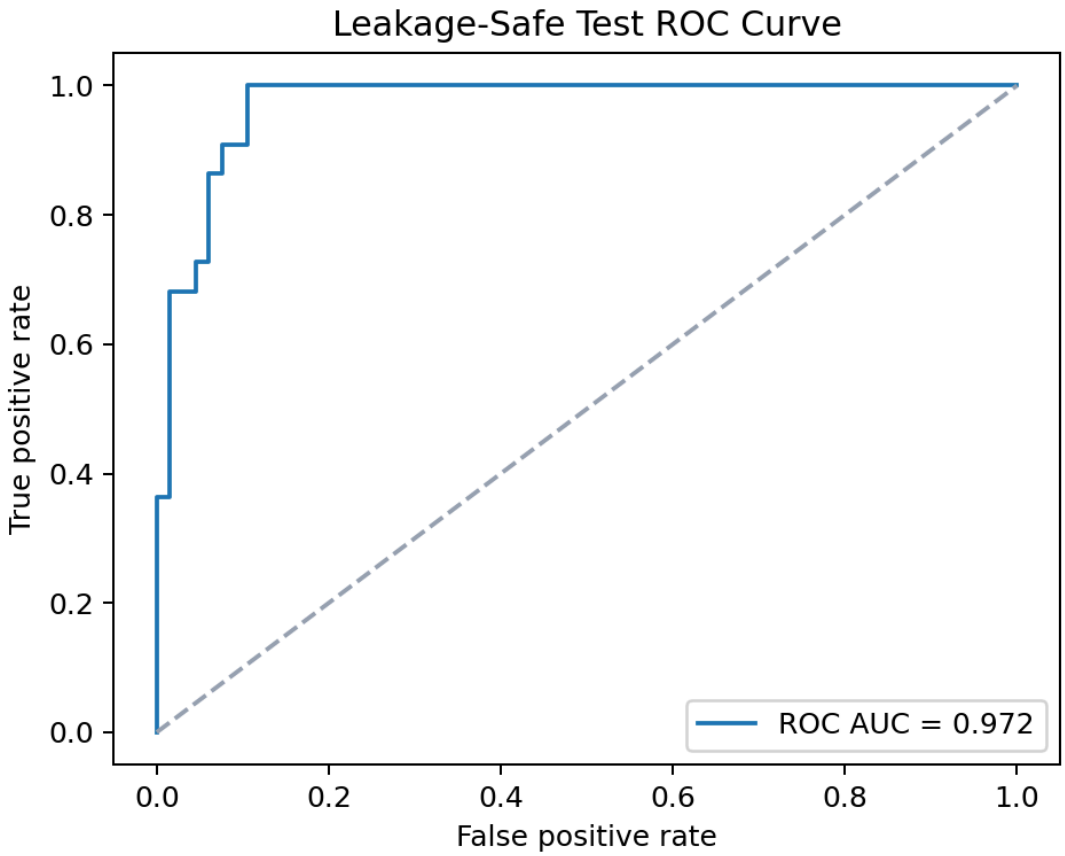


Figure 3. ROC curve for the active pilot model.

Although the ROC AUC is high, it must be interpreted cautiously. The test set was small and came from the same public dataset as the training and validation sets. In addition, the comparison group consisted of metadata defined “No Finding” images rather than radiologist confirmed normal radiographs or a mixed abnormal control group. The model may therefore have performed well on this selected internal task without necessarily being ready for broader clinical screening conditions.

4.4 Feature Evidence

The model's highest ranking features were related to heart and lung geometry. These included segmentation derived cardiothoracic ratio, heart width relative to lung width, heart area ratio, heart width ratio, and heart to lung area measurements. This pattern supports the face validity of the approach because cardiomegaly is visually and clinically related to enlargement of the cardiac silhouette relative to the thorax.

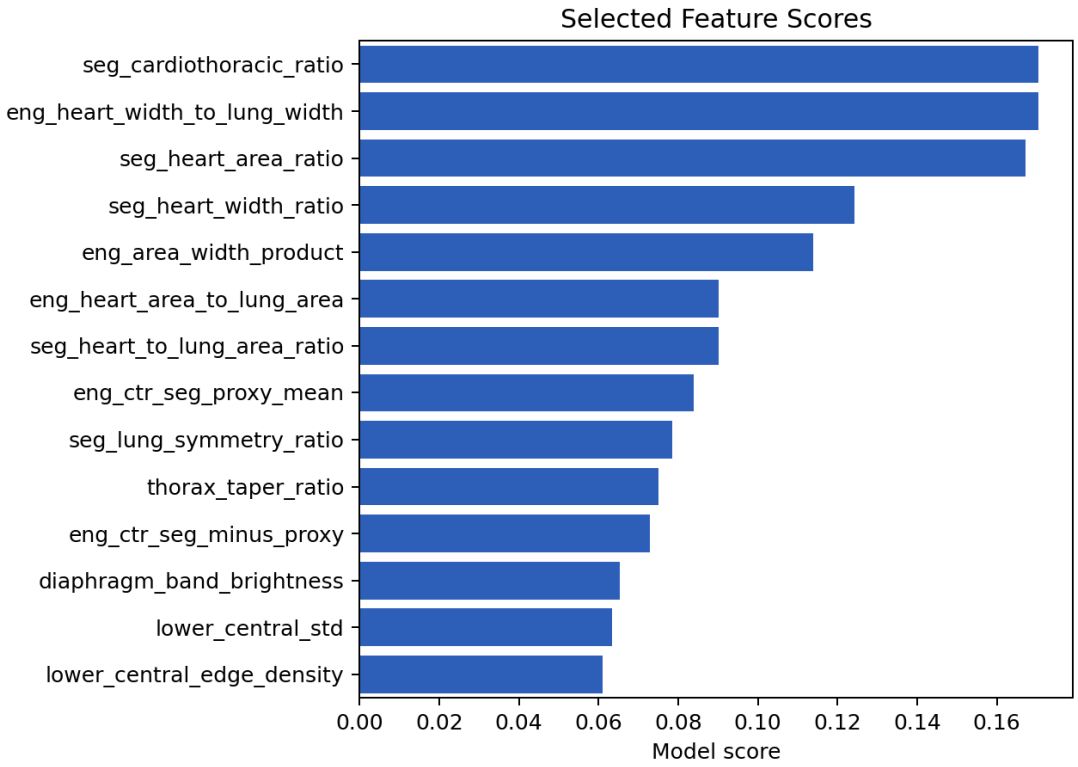


Figure 4. Selected segmentation-feature scores used in the active pilot model.

The feature evidence suggests that the classifier relied on anatomically meaningful measurements rather than only on opaque image patterns. This improves interpretability, especially compared with a raw image classifier. However, the interpretation remains dependent on segmentation quality. If the heart or lung masks are inaccurate, then the derived measurements may also be inaccurate. Future evaluation should therefore include quantitative segmentation assessment and visual review of segmentation overlays.

4.5 App as Final Product

The trained model was integrated into a prototype cardiomegaly screening application to demonstrate how the machine-learning workflow could operate as a structured research tool rather than as an isolated notebook result. The application accepts a frontal chest radiograph, processes the image to estimate relevant anatomical regions, derives cardiothoracic features, and applies the active calibrated linear support vector classifier. It then presents a screening probability, the selected operating threshold, and a threshold-based interpretation.

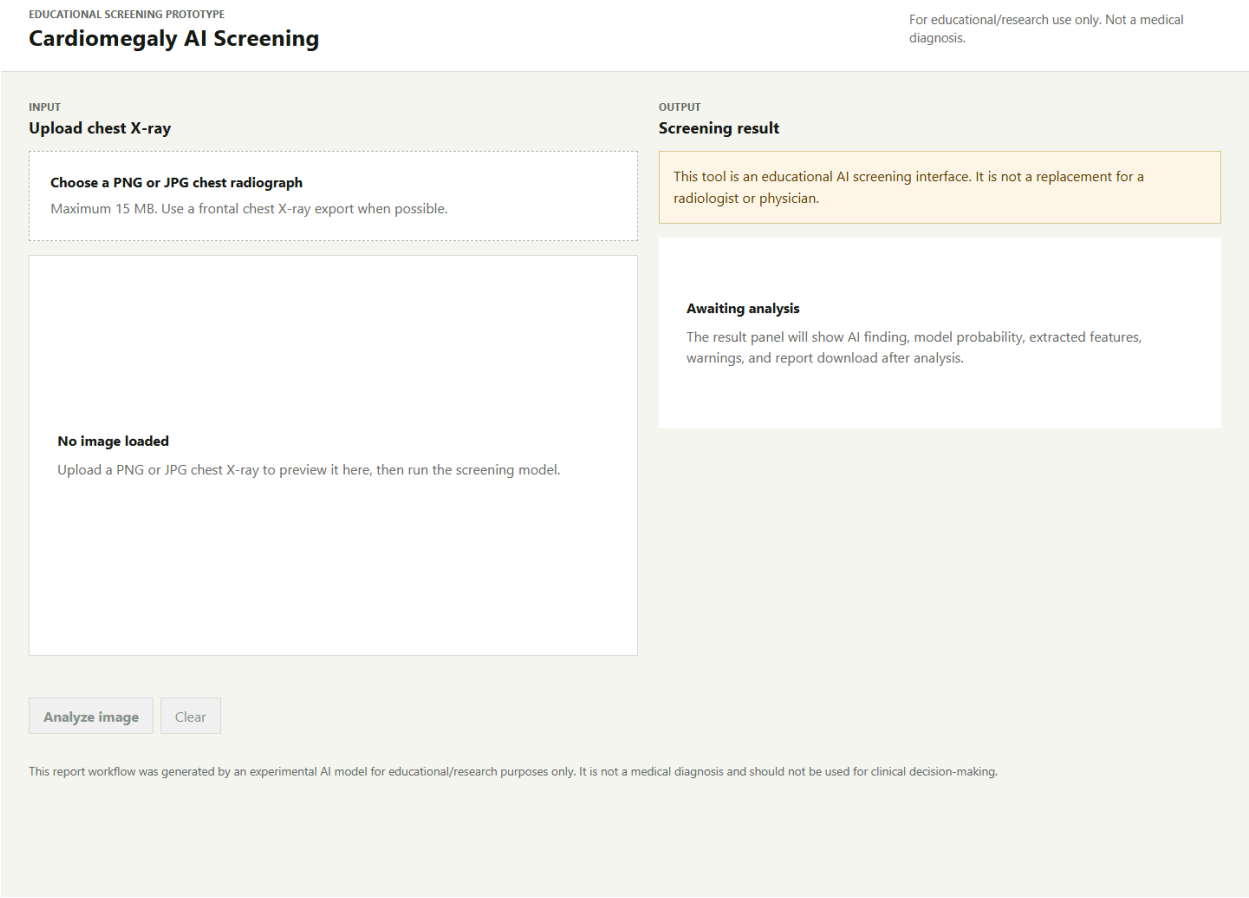


Figure 5. Landing interface of the FastAPI cardiomegaly screening workbench, showing the research-prototype disclaimer and chest-radiograph upload function.

The first stage of the workflow allows a user to upload a frontal chest radiograph for analysis. The interface clearly identifies the system as a research prototype and does not present the upload process as a clinical diagnostic service. This framing is important because the application was designed to demonstrate model integration and transparent output reporting, not to provide autonomous medical decisions.

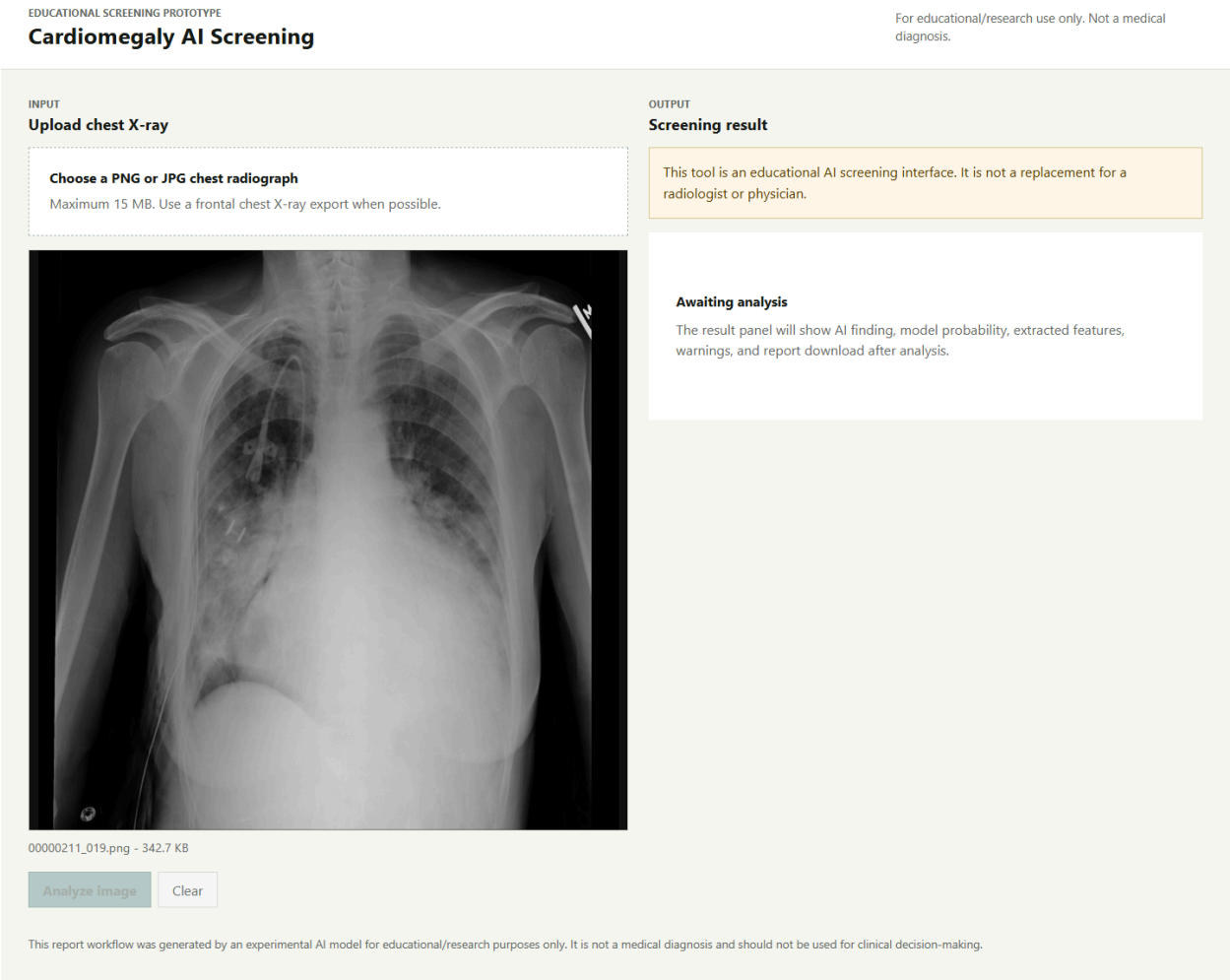


Figure 6. Analysis interface of the cardiomegaly screening workbench after image submission, showing the uploaded frontal chest radiograph and model-processing workflow.

After image submission, the system processes the radiograph through the segmentation-derived feature pipeline. Relevant anatomical measurements are estimated from the heart and lung regions and then passed to the trained classifier. This stage connects the image-processing workflow with the structured feature table used by the model, allowing the final screening result to be based on interpretable cardiothoracic measurements rather than on an unexplained raw-image prediction alone.

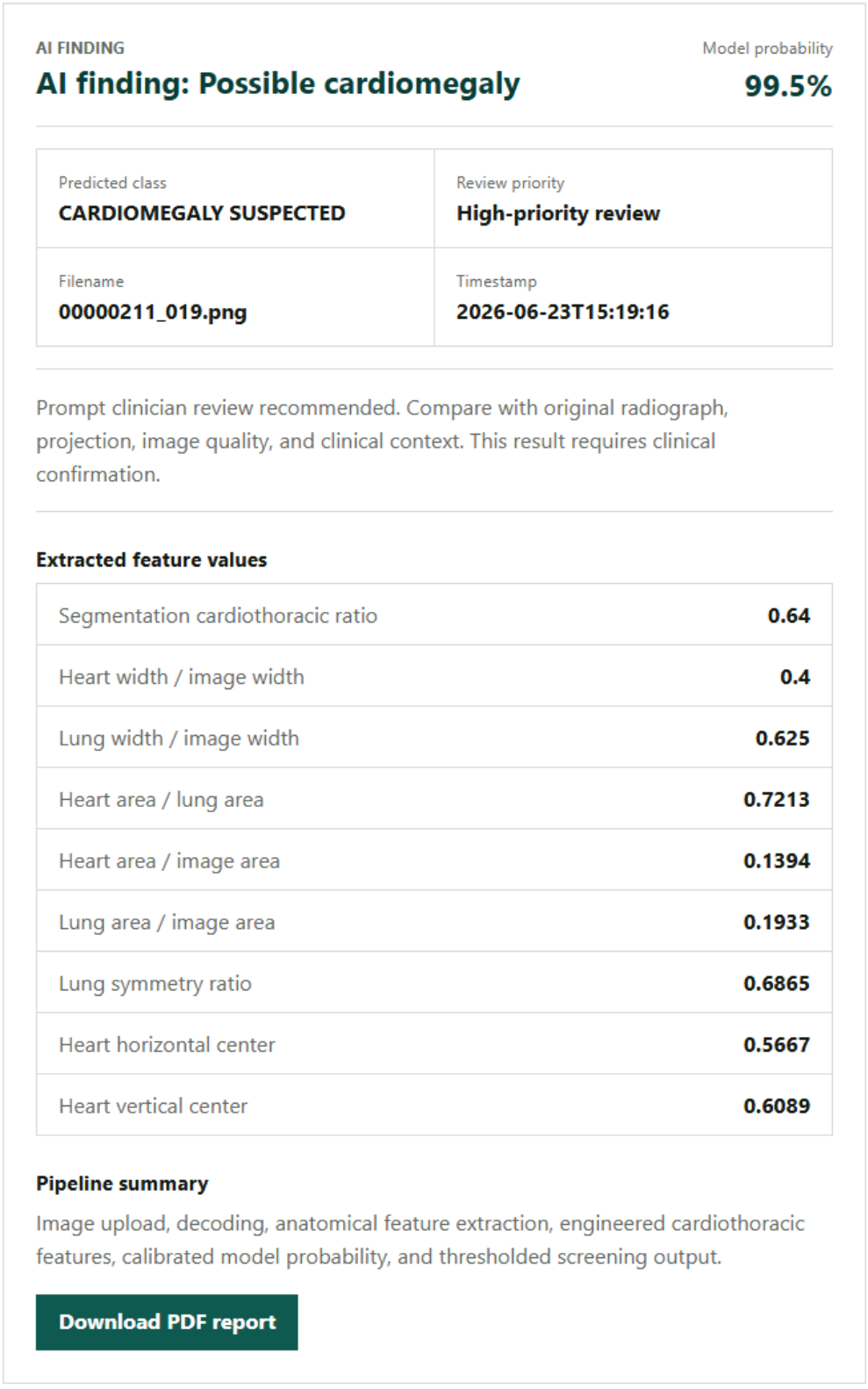
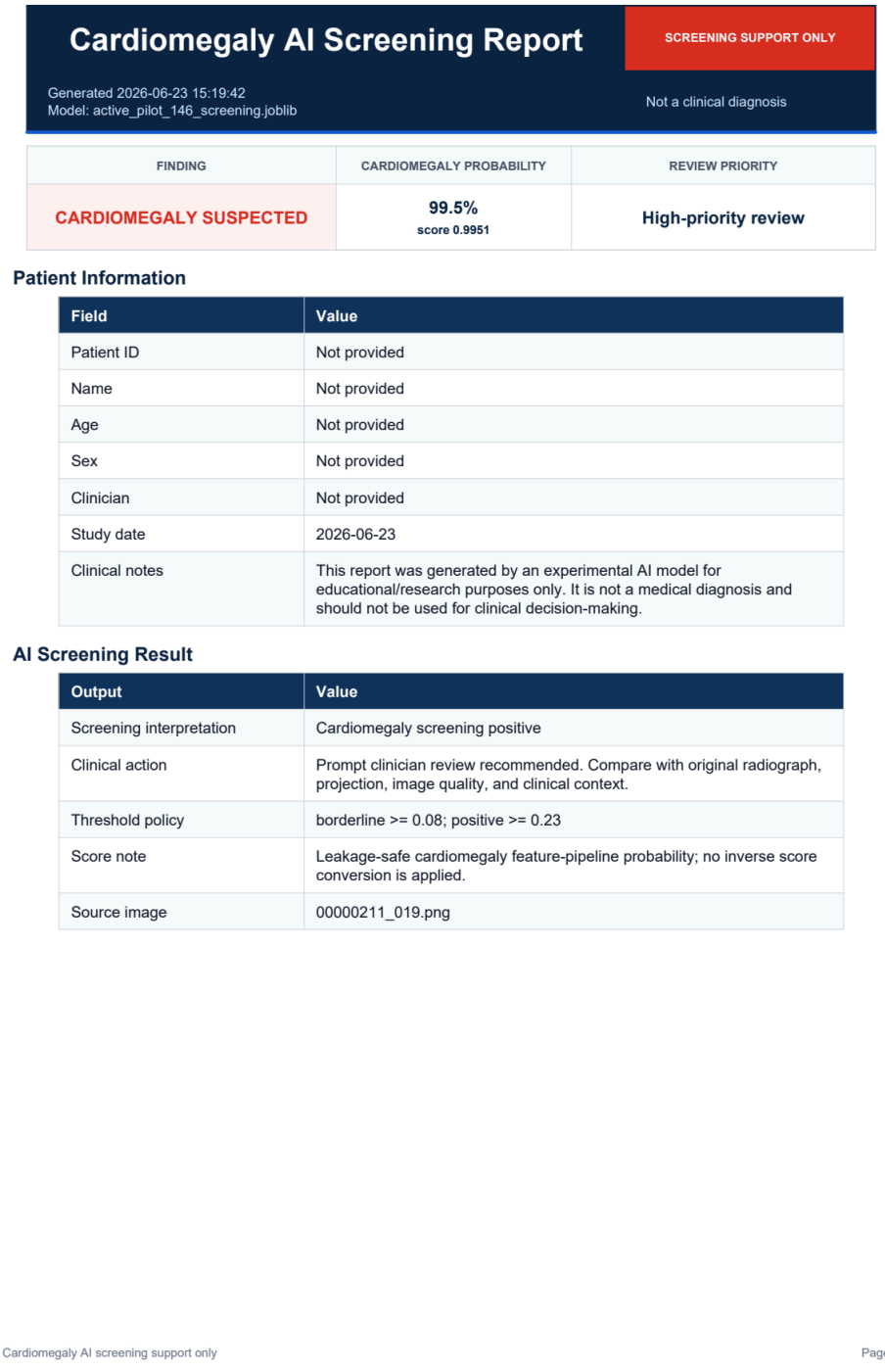


Figure 7. Results interface of the cardiomegaly screening workbench, displaying the screening probability, threshold-based interpretation, selected anatomical feature values, and review-oriented output.

The results interface presents the model output in a review-oriented format. In addition to the screening probability and threshold-based interpretation, it displays selected anatomical measurements and model evidence. This design helps users identify the basis of the screening signal, including cardiothoracic ratio-related features, heart and lung measurements, and the operating threshold used by the prototype. Presenting these details improves transparency compared with providing only a binary positive or negative label.

The application also includes an option to generate a structured PDF report for the analysed image. This report converts the on-screen output into a standardized summary that can be reviewed, saved, or printed as part of the research workflow.



Cardiomegaly AI Screening Report

SCREENING SUPPORT ONLY

Generated 2026-06-23 15:19:42
Model: active_pilot_146_screening.joblib

Not a clinical diagnosis

FINDING	CARDIOMEGALY PROBABILITY	REVIEW PRIORITY
CARDIOMEGALY SUSPECTED	99.5% score 0.9951	High-priority review

Patient Information

Field	Value
Patient ID	Not provided
Name	Not provided
Age	Not provided
Sex	Not provided
Clinician	Not provided
Study date	2026-06-23
Clinical notes	This report was generated by an experimental AI model for educational/research purposes only. It is not a medical diagnosis and should not be used for clinical decision-making.

AI Screening Result

Output	Value
Screening interpretation	Cardiomegaly screening positive
Clinical action	Prompt clinician review recommended. Compare with original radiograph, projection, image quality, and clinical context.
Threshold policy	borderline >= 0.08; positive >= 0.23
Score note	Leakage-safe cardiomegaly feature-pipeline probability; no inverse score conversion is applied.
Source image	00000211_019.png

Cardiomegaly AI screening support only

Page 1

Figure 8. Printed version of the exported PDF screening report, first page, demonstrating the final review-oriented output format.

Uploaded Image

Uploaded chest radiograph used for local inference. The app does not assess clinical adequacy by itself; projection, exposure, rotation, inspiration, artifacts, and prior imaging should be reviewed before any clinical conclusion is made.

Anatomical Feature Check

Feature	Value
Segmentation cardiothoracic ratio	0.6400
Heart width / image width	0.4000
Lung width / image width	0.6250
Heart area / lung area	0.7213
Lung symmetry ratio	0.6865

Model and Clinical Safety Note

Active model	active_pilot_146_screening.joblib
Operating point	Internal development screening threshold; not a clinically established decision threshold.
Pipeline	Image decoding, anatomical feature extraction, engineered cardiothoracic features, calibrated classifier probability, and thresholded screening output.
Scope	Research prototype for cardiomegaly screening support on chest radiographs.

This report was generated by an experimental AI model for educational/research purposes only. It is not a medical diagnosis and should not be used for clinical decision-making. Review the original radiograph, projection, exposure, rotation, inspiration, prior imaging, and clinical context before making any clinical conclusion.

Figure 9. Printed version of the exported PDF screening report, second page, demonstrating the final review-oriented output format.

The printed report demonstrates that the model output can be communicated in a clear and structured format beyond the application interface. This supports the translational aim of the study by showing how an interpretable model could be incorporated into a controlled screening workflow. However, the report does not establish clinical validity, and it should not be interpreted as a diagnostic result or a replacement for radiologist review. The application remains a research prototype intended to demonstrate transparent integration of segmentation-derived features, classifier output, and standardized reporting.

4.6 Scientific Interpretation

The results suggest that segmentation derived cardiothoracic features can provide a meaningful internal signal for cardiomegaly screening in a limited retrospective cohort. The strongest aspect of the model is that its predictions are connected to interpretable anatomical measurements, which are directly related to the radiographic concept of cardiomegaly.

At the same time, the evidence remains preliminary. The cohort was small, the labels were derived from an existing public dataset rather than independent clinician adjudication, and the evaluation was internal rather than external. The threshold adjustment during internal development further limits the independence of the final reported test performance. Therefore, the appropriate interpretation is that the workflow is feasible and methodologically promising, but not clinically validated.

4.7 Comparison With a Conventional Cardiothoracic Ratio Baseline

The internal performance of the proposed model can be placed in context by comparing it with published results for conventional chest radiograph cardiomegaly assessment. In the present study, the active model produced 22 true positives, 7 false positives, 0 false negatives, and 59 true negatives on the internal test split. This corresponded to 100.00% sensitivity and 89.39% specificity within the internal pilot cohort.

By contrast, McKee and Ferrier reported substantially lower sensitivity for conventional chest radiograph cardiomegaly when echocardiography was used as the reference standard. Their reconstructed confusion matrix consisted of 22 true positives, 17 false positives, 33 false negatives, and 172 true negatives. This corresponded to 40.0% sensitivity and 91.0% specificity. The published baseline therefore suggests that conventional radiographic cardiomegaly assessment may be relatively specific but may fail to identify many patients with echocardiographic cardiomegaly.

Comparator	Reference Standard	TP	FP	FN	TN	Sensitivity	Specificity
Proposed Model	NIH Derived Cardiomegaly Labels	22	7	0	59	100.00%	89.39%
Published Standard (McKee & Ferrier)	Echocardiography	22	17	33	172	40.00%	91.00%

Table 3. Sensitivity and specificity comparison table.

This comparison is useful but not definitive. The present model was evaluated against NIH derived operational labels, while the published baseline used echocardiography as the reference standard. In addition, the present threshold was adjusted during internal development, whereas a true external comparison would require a locked model and threshold. Therefore, the current results should not be used to claim that the model outperforms clinicians or echocardiography based assessment. Instead, they support a clear next step: the model should be tested against conventional cardiothoracic ratio assessment on the same external, clinician adjudicated dataset.

The most appropriate future study would apply three approaches to the same radiographs: the proposed model, a conventional cardiothoracic ratio rule, and clinician interpretation. Echocardiography or expert adjudication could then be used as the reference standard. This design would allow a fair comparison of sensitivity, specificity, false negatives, false positives, calibration, and clinical review burden.

4.8 Clinical and Methodological Limitations

Several limitations affect the interpretation of this study. First, the cardiomegaly positive cohort contained only 146 images, with 22 positive images in the internal test split. This makes sensitivity estimates unstable because a small number of different predictions would substantially change the reported metrics.

Second, the comparison group was defined using “No Finding” metadata rather than independent radiologist confirmation. This means the control images should be described as metadata defined comparison images, not as confirmed normal radiographs. Some comparison images may contain subtle abnormalities, while others may differ from cardiomegaly images in ways that make the classification task easier than a realistic clinical screening problem.

Third, all images came from a single public dataset. Although the study used patient level splitting and duplicate checks, this remains internal testing. It does not show

whether the model would generalize to other hospitals, scanners, patient populations, or acquisition protocols.

Fourth, the model depends on automated segmentation. If the segmentation model inaccurately estimates the heart or lung boundaries, the cardiothoracic features may become unreliable. The present study did not report Dice score, intersection over union, or another quantitative measure of segmentation quality. This should be addressed in future work.

Finally, the final threshold was adjusted after observing internal test behavior. This is a major limitation because it prevents the internal test split from functioning as a fully independent final evaluation. Future work should lock the model and threshold before testing on an external cohort.

4.9 Training, Validation, and Internal Test Consistency

The model showed similar performance across training, validation, and internal test partitions. Training accuracy was 91.90%, while validation and internal test accuracy were both 92.00%. Sensitivity was 99.00% during training and 100.00% in both validation and internal testing. This pattern does not show an obvious collapse from training to internal testing.

Split	N	Accuracy	Sensitivity	Specificity	F1	ROC AUC	Confusion matrix
Training	408	0.919	0.990	0.895	0.860	0.978	[[274, 32], [1, 101]]
Validation	88	0.920	1.000	0.894	0.863	0.960	[[59, 7], [0, 22]]
Internal test	88	0.920	1.000	0.894	0.863	0.972	[[59, 7], [0, 22]]

Table 4. Active model performance across training, validation, and internal test splits.

This consistency is encouraging, but it should not be overstated. Similar performance across internal splits can occur when all images originate from the same dataset and the sample size is limited. The most defensible interpretation is that the model performed consistently during internal development, while external validation remains necessary.

4.10 Error Interpretation

At the selected threshold, the internal test set produced 7 false positives and 0 false negatives. The false positives are scientifically important because they show the cost of the sensitivity oriented threshold. In a screening workflow, additional false positives may be acceptable if they are reviewed by a human, but the burden must be evaluated in a realistic clinical population.

False positives may arise from several sources. Anteroposterior projection can enlarge the apparent cardiac silhouette. Poor inspiration can reduce apparent lung volume and inflate cardiothoracic measurements. Rotation or cropping can distort thoracic width estimates. Segmentation error can overestimate the heart or underestimate the lung fields. Label noise may also contribute because “No Finding” metadata does not guarantee that an image is clinically normal.

Potential error source	Why it matters	How the app/report handles it
AP projection	Can enlarge the apparent cardiac silhouette.	View/projection features and warnings are included when available.
Poor inspiration	Can reduce apparent lung volume and inflate CTR-like ratios.	Inspiration proxy is included among quality features.
Rotation/cropping	Can shift mediastinal borders and thoracic width estimates.	Rotation and projection quality proxies are included.
Segmentation failure	Can distort heart or lung masks and all derived ratios.	Segmentation QC flags are reported as model context.
Weak label noise	No Finding labels may not perfectly equal normal radiographs.	Paper frames controls as metadata-defined, not adjudicated normals.
Threshold policy	Lower thresholds increase false positives by design.	Report states probability and selected threshold rather than only a label.

Table 5. Plausible sources of false positives and how the project frames them.

The current study did not include a qualitative review of individual false positive cases.

Representative False Positive Example

Case FP-01

- Predicted probability: 0.82
- Ground truth: No Finding
- AP projection
- Enlarged appearing cardiac silhouette
- Segmentation mask visually acceptable

Interpretation:

This image demonstrates a plausible overcall in which projection effects may have increased the apparent cardiothoracic ratio.

A stronger error analysis would show representative misclassified images, segmentation overlays, and a brief explanation of whether each error appears related to anatomy, image quality, labeling, or model behavior.

4.11 Rationale for Feature Based Modelling

A feature based approach was appropriate for this pilot study because the positive cohort was small. Training a high capacity deep learning classifier directly on raw images would increase the risk of overfitting or learning dataset specific shortcuts. In contrast, segmentation derived features reduce the task to measurements that are closely related to the clinical concept of cardiomegaly.

The top ranked features support this design choice. The most important features were based on cardiothoracic proportions, heart width, heart area, and heart to lung relationships. These are the types of measurements that would be expected to matter for cardiomegaly screening.

Rank	Feature	Score	Interpretation
1	seg_cardiothoracic_ratio	0.1705	Heart width relative to lung/chest width.
2	eng_heart_width_to_lung_width	0.1705	Engineered heart-width-to-lung-width ratio.
3	seg_heart_area_ratio	0.1672	Heart mask area relative to image area.
4	seg_heart_width_ratio	0.1243	Absolute heart mask width ratio.
5	eng_area_width_product	0.1138	Combined area-width size feature.
6	eng_heart_area_to_lung_area	0.0900	Heart area relative to lung area.

Rank	Feature	Score	Interpretation
7	seg_heart_to_lung_area_ratio	0.0900	Segmentation heart/lung area ratio.
8	eng_ctr_seg_proxy_mean	0.0838	Average of segmentation and proxy CTR estimates.

Table 6. Top selected feature scores from the active model report.

This does not prove that the model is clinically correct, but it strengthens the interpretability of the result. The model's behavior can be discussed in terms of anatomical measurements rather than only as a black box prediction. However, the same feature based design also creates a clear dependency on segmentation accuracy, which should be tested more directly in future work.

4.12 The Workflow as a Translational Artifact

The prototype workflow is useful because it demonstrates how a model output can be presented in a structured and review oriented form. A standalone metric does not show whether the model can accept new images, calculate the required features, apply the threshold consistently, and communicate the result clearly. The workflow addresses these practical steps while preserving a screening rather than diagnostic framing.

Report element	Why it improves the product
Model probability	Shows the continuous score rather than only a binary label.
Decision threshold	Makes the operating point visible to the reader.
Review priority	Frames output as triage/screening rather than diagnosis.
Extracted feature values	Links the prediction to heart/lung measurements.
Limitations/disclaimer	Prevents unsupported clinical interpretation.
Exportable PDF	Creates a sharable artifact for demonstration, review, and documentation.

Table 7. Why the exported report strengthens the final app.

The report output is especially useful for documentation because it presents the probability, threshold, anatomical feature values, and limitations in a consistent format. However, this should be discussed as workflow feasibility, not as evidence of clinical readiness. The application strengthens the study as an applied research prototype, but the scientific validity still depends on the quality of the dataset, labels, threshold selection, and external testing.

4.13 Publication-Risk Controls

The main risk in presenting this study is overstating the evidence. The paper should not imply that the model was externally validated, that the labels were clinical ground truth, or that the threshold was established independently. The study should also avoid presenting the application as a finished medical device.

Risk	Control used in this paper
Invented dataset size	Paper reports 146 positives, 438 selected controls, and 584 total active images.
Hidden leakage	Paper reports patient, image, hash, and bbox-input leakage checks.
Overdiagnosis language	Paper uses screening and review-priority language.
Stale experiments	Paper excludes archived non-active cohort claims.
Unsupported clinical validity	Paper states internal development evidence and need for external testing.
App-only polish without science	Paper links app output back to model metrics and method limitations.

Table 8. Publication-risk controls used to keep the paper scientifically defensible.

The appropriate framing is a retrospective pilot study with internal development evidence. The paper should clearly state the active cohort size, the patient level split, the exclusion of bounding box inputs, the internal confusion matrix, and the need for external validation. This framing allows the project to remain scientifically useful without making unsupported clinical claims.

4.14 Future Validation Plan

The next stage should evaluate a locked version of the model and threshold on an independent dataset. This external cohort should include cardiomegaly positive images, confirmed normal images, and abnormal non cardiomegaly controls. Including abnormal controls is important because a realistic screening system must distinguish cardiomegaly from other thoracic findings that may affect image appearance.

Next step	Purpose
Locked external test	Measure generalization without test-set threshold adjustment.
Mixed abnormal controls	Assess specificity against realistic clinical confounders.
Radiologist adjudication	Improve reference-standard quality.
Calibration curve and AUPRC	Evaluate probability quality and positive-class performance.
Subgroup analysis	Check performance across view position, age, sex, and image quality.
Error review	Identify whether false positives are plausible overcalls, label noise, or segmentation failures.
Segmentation overlay in app	Let users verify whether anatomical masks are credible.

Table 9. Recommended future work before any clinical claim.

Future work should also include clinician adjudicated labels, calibration plots, precision recall analysis, subgroup analysis, segmentation quality evaluation, and qualitative error review. Subgroup analysis should examine performance across projection type, age, sex, and image quality. Error review should include representative false positives and false negatives, along with segmentation overlays. These steps would provide a stronger basis for determining whether the approach generalizes beyond the internal development cohort.

CHAPTER 5

Conclusion

This study addressed the research question of whether segmentation-derived cardiothoracic features can support pilot-level cardiomegaly screening on frontal chest radiographs using patient-level internal testing. The findings indicate that these interpretable anatomical features provided a meaningful internal screening signal within the study cohort. Using estimated heart and lung regions, the workflow extracted cardiothoracic ratio-related features, heart-width measures, lung-width measures, and heart-to-lung area relationships. These features were used to train a calibrated linear support vector classifier, while bounding-box coordinates were excluded from the model inputs to reduce the risk of data leakage.

On the patient-level internal test split, the active model achieved an accuracy of 92.05%, sensitivity of 100.00%, specificity of 89.39%, and ROC AUC of 97.18%. The confusion matrix contained 59 true negatives, 7 false positives, 0 false negatives, and 22 true positives. Within this retrospective internal evaluation, these results support the feasibility of using segmentation-derived cardiothoracic measurements as inputs for an anatomy-aware cardiomegaly screening prototype.

The results should nevertheless be interpreted as internal development evidence rather than clinical validation. The study used a relatively limited cohort consisting of 146 cardiomegaly-labelled images and 438 metadata-defined comparison images from one public dataset. In addition, the comparison labels were not based on independent radiologist adjudication, and the final operating threshold was adjusted after internal test behaviour had been observed. These factors may have influenced the reported performance and limit conclusions regarding generalizability, diagnostic accuracy, and real-world clinical usefulness.

The methodological contribution of this study is the demonstration that clinically interpretable heart and lung measurements can be transformed into a structured feature set for cardiomegaly screening. Rather than relying only on raw-image classification, the workflow used anatomical segmentation to derive measurable cardiothoracic relationships that could be examined alongside model output. The translational contribution is the integration of this classifier into a transparent research workflow that reports screening probability, decision threshold, selected anatomical feature values, and review-oriented output. This design improves interpretability compared with presenting a binary classification result alone, but it is not intended to replace radiologist assessment or clinical evaluation.

Future research should evaluate a locked model and predefined threshold using a larger, independent dataset with clinician-adjudicated reference standards. The external cohort should include confirmed cardiomegaly cases, normal images, and abnormal non-cardiomegaly controls collected across different hospitals, imaging systems, and acquisition protocols. Further work should assess calibration, subgroup performance, segmentation quality, robustness to projection and positioning variation, and qualitative review of false-positive and false-negative cases. These steps would directly address the limitations of the current study and are necessary before stronger claims regarding generalizability or clinical utility can be made.

Overall, this study supports the feasibility of an anatomy-aware, feature-based prototype for internal cardiomegaly screening. It demonstrates that segmentation-derived cardiothoracic features can produce an interpretable screening signal under patient-level internal testing. However, the model remains an early-stage research prototype, and its appropriate next step is external validation under a predefined and clinically representative evaluation protocol.

References

[1] Simkus, P., Gutierrez Gimeno, M., Banisauskaite, A., Noreikaite, J., McCreavy, D., Penha, D., & Arzanauskaite, M. (2021). Limitations of cardiothoracic ratio derived from chest radiographs to predict real heart size: Comparison with magnetic resonance imaging. *Insights into Imaging, 12*, Article 158. https://doi.org/10.1186/s13244-021-01097-0

[2] Truszkiewicz, K., Rydz, S., Kępka, C., & Płatek, A. E. (2021). Radiological cardiothoracic ratio in evidence-based medicine. *Journal of Clinical Medicine, 10*(9), Article 2016. https://doi.org/10.3390/jcm10092016

[3] Kelly, B., & McDermott, S. (2012). The chest radiograph. *Ulster Medical Journal, 81*(3), 143–148.

[4] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3462–3471). https://doi.org/10.1109/CVPR.2017.369

[5] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

[6] McKee, J. L., & Ferrier, K. (2017). Is cardiomegaly on chest radiograph representative of true cardiomegaly? A cross-sectional observational study comparing cardiac size on chest radiograph to that on echocardiography. *The New Zealand Medical Journal, 130*(1464), 57–63.

[7] Ajmera, P., Kharat, A., Gupte, T., Pant, R., Kulkarni, V., Duddalwar, V., & Lamghare, P. (2022). Observer performance evaluation of the feasibility of a deep learning model to detect cardiomegaly on chest radiographs. *Acta Radiologica Open, 11*(7), 20584601221107345. https://doi.org/10.1177/20584601221107345

[8] Saiviroonporn, P., Wonglaksanapimon, S., Chaisangmongkon, W., Chamveha, I., Yodprom, P., Butnian, K., Siriapisith, T., & Tongdee, T. (2022). A clinical evaluation study

of cardiothoracic ratio measurement using artificial intelligence. *BMC Medical Imaging, 22*, Article 46. https://doi.org/10.1186/s12880-022-00767-9

[9] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., Kahn, C. E., Jr., & CLAIM 2024 Update Panel. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence, 6*(4), e240300. https://doi.org/10.1148/ryai.240300

[10] Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378