# **Detecting Left Ventricular Structural Abnormalities in Chest X-Rays using Deep Learning: A Replication Study**

## **Introduction**

Heart failure often develops silently, with early structural changes that precede overt symptoms. Early recognition of left ventricular (LV) structural disease is clinically valuable because initial symptoms can be mild and nonspecific. Bhave et al. (2024) proposed a deep learning approach to screen standard chest radiographs (CXRs) for two echocardiographic phenotypes of structural heart disease; severe LV hypertrophy (SLVH) and dilated LV (DLV) as well as a composite outcome indicating either condition. In their study, CXRs were paired with echocardiograms within a 12-month window, and a convolutional neural network (DenseNet-121 backbone) was trained to jointly leverage image features and patient demographics (age, sex) to predict continuous LV measurements and subsequent binary outcomes (Bhave et al., 2024). The model achieved high discriminative performance on the internal test set (area under the ROC curve [AUROC] ~0.79–0.80 for SLVH, DLV, and the composite outcome) and maintained similar performance on an external dataset, outperforming radiologists in detecting the composite outcome (Bhave et al., 2024). As a resource for the community, the authors publicly released the CXR-echo dataset (71,589 CXRs from 24,689 patients) to enable replication and further innovation (Bhave et al., 2024).

Our objective in this replication study was to reproduce the model of Bhave et al. using the publicly available dataset and to compare the results of our independent implementation with the originally reported discrimination and calibration. We aimed to confirm whether similar performance could be achieved under the same data definitions and to identify any implementation differences that might affect model outcomes.

## **Methods**

### **Data and Preprocessing**

We used the publicly released Columbia University Irving Medical Center (CUIMC) CXR-echo dataset (as described by Bhave et al., 2024) comprising 71,589 frontal chest X-ray images from 24,689 patients. Each CXR had at least one transthoracic echocardiogram within ±12 months, which provided ground truth measurements of LV structure. From the echocardiography reports, we extracted three continuous measurements: interventricular septal thickness in diastole (IVSd), LV posterior wall thickness in diastole (LVPWd), and LV internal diameter in diastole (LVIDd). These measurements were binarized into diagnostic labels for SLVH and DLV using the same sex-specific threshold criteria as the original study **(Bhave et al., 2024)**. SLVH was labeled for men if IVSd > 1.5 cm or LVPWd > 1.5 cm (for women, > 1.4 cm), and DLV was labeled for men if LVIDd > 5.9 cm (for women, > 5.3 cm). A composite label was assigned if either SLVH or DLV criteria were met. These threshold definitions correspond to guideline criteria for severe LV hypertrophy and dilation and mirror those used by Bhave et al. in the original study.

To prevent information leakage between training and testing, we performed a patient-level split of the data. All CXRs from a given patient were confined to a single set. We randomly split the patients into training (90%), validation (5%), and test (5%) subsets. This yielded an internal test set of approximately 3,571 CXRs from 1,235 patients (Table 1). Demographic characteristics and label prevalences were compared across the splits to ensure no significant differences. Prior to model training, each CXR was preprocessed following the procedure described by Bhave et al.: images were converted to grayscale if not already, center-cropped to a square region capturing the thorax and resized to 224×224 pixels using bicubic interpolation. We then applied contrast-limited adaptive histogram equalization (CLAHE) for contrast enhancement, matching the original study’s preprocessing. Pixel intensity values were normalized to the standard range used by the DenseNet-121 ImageNet model. During training, mild data augmentation (random flips and shifts) was employed to improve generalizability, consistent with the augmentation strategy noted in the original implementation.

### **Model Architecture and Training**

We recreated the deep learning model as described by **Bhave et al. (2024)**, with a few modifications in framework but not in overall design. The original model was implemented in PyTorch; our replication used an equivalent architecture in Keras/TensorFlow. Specifically, we used a DenseNet-121 convolutional neural network (pretrained on ImageNet) as the CXR feature extractor, taking the 224×224 preprocessed image as input. The DenseNet backbone’s output (global average pooled feature map) was concatenated with two demographic inputs: patient age (normalized) and sex (encoded as a one-hot vector for female/male). The combined feature vector was fed into *multitask regression heads* to predict the three continuous echocardiographic measures (IVSd, LVPWd, LVIDd) simultaneously. This multitasks design mirrors the original approach, which used the intermediate regression outputs to inform the final classification. In our implementation, each regression head produced both a mean and variance (uncertainty) estimate for its target, and we optimized a Gaussian negative log-likelihood loss for each continuous output. The model was trained end-to-end to minimize the joint loss of all three regression tasks.

For training, we adopted a two-phase strategy like Bhave et al.’s description: an initial “warm-up” phase and a subsequent fine-tuning phase. In the warm-up, we froze the DenseNet-121 backbone weights and trained only the new regression head and demographic fusion layers for 10 epochs, allowing the added layers to learn without disturbing the pretrained imaging features. Then we fine-tuned the model for an additional 40 epochs, unfreezing the last few DenseNet blocks (the extent of unfreezing matched the original paper’s approach) and training with a lower learning rate (1×10^–5) to refine the image feature extractor on this specific task. We used the Adam optimizer for training. The batch size was 32 images. During fine-tuning, the same augmentation (random horizontal flips and small rotations/translations) was applied as in the warm-up. Throughout training, model performance on the validation set was monitored, and the epoch with the best validation loss was used for final evaluation to avoid overfitting.

After training the model to predict the continuous LV measures, we derived the probabilities for each binary outcome. Rather than simply thresholding the predicted values (which was the conceptual approach in the original study to determine SLVH/DLV presence), we implemented a calibration step to optimize classification performance. Using the held-out validation set, we trained a simple logistic regression model to map the three continuous outputs (IVSd, LVPWd, LVIDd predictions) to each binary label. This served as a *learned combiner* that weighted the contributions of each echo measurement to the final probability of SLVH, DLV, or the composite. We then applied isotonic regression on the validation outputs to calibrate the probabilities to the observed label frequencies. The result was a calibrated model that outputs well-conditioned probabilities for each outcome given an input CXR. Finally, we evaluated the calibrated model on the independent test set.

### **Evaluation Metrics**

We assessed model discrimination using receiver operating characteristic (ROC) curves and precision–recall (PR) curves for each of the three outcomes (SLVH, DLV, composite). The primary metrics were the area under the ROC curve (AUROC) and the average precision (AP, i.e. area under the PR curve). To quantify uncertainty in these performance metrics, we computed 95% confidence intervals (CIs) via bootstrap resampling. For each outcome, we performed 1,000 bootstrap replicates of the test-set predictions, calculating AUROC and AP for each replicate. The 2.5th and 97.5th percentiles of these bootstrap distributions were taken as the 95% CI bounds. Model calibration was examined by computing the Brier score for each outcome (mean squared error between predicted probability and the true label, lower values indicating better calibration) and by inspecting reliability curves (not shown). All results from our replication were then compared with the originally published results for the internal test set of the CUIMC dataset (Bhave et al., 2024).

## **Results**

### **Cohort Characteristics**

A total of 71,589 CXRs from 24,689 patients were included in the study after applying the inclusion criteria (CXRs with an echocardiogram within one year). The training set contained 90% of the patients (≈22,200 patients, 64,467 images), while the validation and test sets each contained 5% of patients (~1,234–1,235 patients, 3,551–3,571 images) (Table 1). The mean age of patients was approximately 62 years in each subset, with a standard deviation of about 16 years. The age distribution was similar across training, validation, and test splits (about 38% of patients were under 59 years, 27% in their 60s, 21% in their 70s, and ~13–14% aged 80 or above in each subset). Each subset had roughly half of patients being female (overall 52.1% female). Importantly, the prevalence of the target conditions was low in all splits, reflecting the class imbalance inherent in this clinical scenario. In the test set, 8.8% of CXRs had evidence of SLVH and 10.0% had DLV, with 16.3% positive for either condition (composite outcome). These prevalences were closely matched between the training and test sets (composite prevalence ~17% in both).

***Table 1.*** *Patient characteristics and label prevalence in the CUIMC CXR-echo dataset, for the training, validation, and test splits used in this replication study.*

| Section | Metric | Train | Val | Test | Overall |
| --- | --- | --- | --- | --- | --- |
| **Patient characteristics** | CXRs, n | 64,467 | 3,551 | 3,571 | 71,589 |
|  | Patients, n | 22,220 | 1,234 | 1,235 | 24,689 |
|  | Age, years (mean ± SD) | 62.2 ± 16.1 | 61.7 ± 16.3 | 62.0 ± 15.9 | 62.1 ± 16.1 |
|  | Age < 59, n (%) | 24,525 (38.0%) | 1,393 (39.2%) | 1,458 (40.8%) | 27,376 (38.2%) |
|  | Age 60–69, n (%) | 17,535 (27.2%) | 926 (26.1%) | 898 (25.1%) | 19,359 (27.0%) |
|  | Age 70–79, n (%) | 13,551 (21.0%) | 774 (21.8%) | 717 (20.1%) | 15,042 (21.0%) |
|  | Age ≥ 80, n (%) | 8,856 (13.7%) | 458 (12.9%) | 498 (13.9%) | 9,812 (13.7%) |
|  | Female sex, n (%) | 33,486 (51.9%) | 1,845 (52.0%) | 1,958 (54.8%) | 37,289 (52.1%) |
| **Echo measures** | IVSd (cm, mean ± SD) | 1.1 ± 0.3 | 1.1 ± 0.2 | 1.1 ± 0.2 | 1.1 ± 0.3 |
|  | LVPWd (cm, mean ± SD) | 1.1 ± 0.2 | 1.1 ± 0.2 | 1.1 ± 0.2 | 1.1 ± 0.2 |
|  | LVIDd (cm, mean ± SD) | 4.6 ± 0.7 | 4.6 ± 0.6 | 4.6 ± 0.7 | 4.6 ± 0.7 |
| **Outcome prevalence** | SLVH, n (%) | 6,467 (10.0%) | 317 (8.9%) | 313 (8.8%) | 7,097 (9.9%) |
|  | DLV, n (%) | 6,250 (9.7%) | 336 (9.5%) | 357 (10.0%) | 6,943 (9.7%) |
|  | Composite (SLVH or DLV), n (%) | 11,167 (17.3%) | 574 (16.2%) | 581 (16.3%) | 12,322 (17.2%) |
|  |  |  |  |  |  |

*Interpretation of Table 1:* The training, validation, and test sets are well-matched in terms of patient demographics and outcome prevalence. The overall cohort is middle-aged to elderly (mean age ~62 years) with a slight majority of patients being female. The distribution across age brackets is uniform between the splits, and the proportion of females is ~52% in both training and test sets. The average echocardiographic measurements (IVSd, LVPWd, LVIDd) are identical across subsets (mean septal and posterior wall thickness ~1.1 cm, mean LV end-diastolic diameter ~4.6 cm), indicating no appreciable differences in baseline cardiac structure between the groups. Approximately 10% of patients meet criteria for severe LV hypertrophy or dilation individually, and 17% have either abnormality. The composite outcome is more common than each individual condition, as expected, but still present in only roughly one in six patients. The similarity of these statistics across training and test sets suggests that the random splitting did not introduce any significant sampling bias, allowing a fair evaluation of model generalizability on the internal test data.

### **Discrimination Performance**

On the held-out CUIMC test set, our replicated model demonstrated good discrimination for all three outcomes, though performance was slightly below that reported in the original study.

A graph with a line

AI-generated content may be incorrect.

**Figure 1.** Receiver operating characteristic curves for the deep learning model’s performance in detecting LV structural abnormalities on the internal test set (CUIMC), with bootstrap-derived 95% confidence intervals. Each curve represents one outcome: the composite outcome (either SLVH or DLV, blue), DLV alone (orange), and SLVH alone (green). The diagonal gray line is the line of no-discrimination (AUROC = 0.50). The model achieves an area under the ROC curve (AUROC) of 0.75 for the composite outcome, 0.77 for DLV, and 0.73 for SLVH. The shaded regions (and values in the legend) indicate the 95% confidence intervals for each AUROC, derived from 1,000 bootstrap samples. All three curves lie well above the diagonal, indicating that the model can discriminate patients with structural abnormalities from those without. Notably, the DLV curve is slightly higher than the SLVH curve, suggesting the model was more effective at detecting LV dilation than LV hypertrophy. The composite outcome, which combines both, shows intermediate performance. The overlap of confidence intervals implies these differences are modest, and overall, the model demonstrates **moderate discriminative power** for all outcomes.

A graph of a graph

AI-generated content may be incorrect.

Following the ROC analysis, **Figure 2** presents the precision–recall (PR) curves for the same three outcomes, which more directly highlight performance on the positive class. The PR curves are especially informative given the class imbalance (relatively low prevalence of SLVH/DLV). The model’s precision–recall curve for the composite outcome (blue) yields an average precision of about 0.31, which is substantially higher than the baseline positive rate of ~0.16 for the composite condition. This indicates that the model concentrates a significant fraction of true positives towards the top of its risk predictions. The PR curves for DLV (orange) and SLVH (green) have lower average precision (~0.19 each), reflecting the lower prevalence (~9–10%) of each condition and the challenge of identifying these cases. Nonetheless, both curves stay above their respective baseline precision levels, confirming that the model’s predictions carry useful signal. The initial steep portion of each PR curve suggests that at high recall levels (approaching all positives found), precision declines but still remains above random, which is expected in a medical screening context. In summary, the PR analysis shows that the model can identify a subset of patients with structural heart disease with reasonable precision, though many false positives occur once recall is pushed higher (a typical trade-off in imbalanced classification scenarios).

### **Comparison to Original Study**

Our replication results can be directly compared to the internal test performance reported by **Bhave et al. (2024)** on the same CUIMC dataset. The original model had slightly higher discrimination: they reported AUROC values of 0.80 for DLV, 0.79 for SLVH, and 0.80 for the composite outcome (with overlapping CIs in the range ~0.77–0.84), versus our reproduced AUROCs of 0.766, 0.729, and 0.750, respectively. The absolute differences in AUC are on the order of 0.03–0.05 (3–5 percentage points). In terms of precision–recall, the original study’s composite outcome had an AP of approximately 0.53 on the internal test (benefiting from a somewhat higher prevalence in their test distribution or thresholding on echo measurements), compared to our composite AP of 0.31. Overall, our model’s rank-ordering of cases by risk is similar but slightly less sharp than in the original model.

It is worth noting that the original study also validated their model on an external dataset of 8,003 CXRs from a different hospital and found broadly similar performance (AUROCs ~0.67–0.78) albeit with some drop for the SLVH label (Bhave et al., 2024). They further showed that their model could outperform 15 board-certified radiologists in detecting the composite outcome at a fixed specificity of 73% (sensitivity 71% for the model vs. 66% for the radiologist consensus) (Bhave et al., 2024). These aspects underscore the clinical relevance of the approach. Our replication focused only on the internal dataset and did not include an external validation or reader study. We therefore report exclusively on the reproducibility of the internal discrimination results.

## **Discussion**

We successfully replicated the core design and methodology of **Bhave et al. (2024)** for CXR-based detection of LV structural abnormalities, and we demonstrated that an independent implementation could achieve **moderately strong discrimination** on the same large clinical dataset. The patient-level data split, sex-specific diagnostic thresholds, and DenseNet-121 image backbone with age/sex feature fusion were all implemented as described in the original paper. The resulting model’s performance (AUROC ~0.73–0.77 across tasks) confirms the feasibility of screening CXRs for echocardiographic evidence of severe hypertrophy or dilation. However, our reproduced results were modestly lower than those originally reported (AUROCs ~0.79–0.80 in Bhave et al., 2024). There are several plausible factors that could explain this performance gap between our replication and the original model such as dataset split and prevalence, training regimen, and augmentation intensity.

Despite these differences, it is heartening that our replication achieved broadly similar results, reinforcing the soundness of the original study’s conclusions. The model’s ability to detect SLVH and DLV from CXRs even with slightly reduced accuracy is remarkable, considering that radiologists do not routinely diagnose these conditions from a chest X-ray alone. The replication process highlights the importance of implementation details: replicating medical AI studies often requires careful attention to seemingly minor factors like preprocessing and calibration, which can meaningfully affect outcomes.

In conclusion, this replication study confirms that deep learning analysis of CXRs can detect certain LV structural abnormalities (hypertrophy and dilation) with a performance approaching that of the original reported model. Small deviations in implementation can lead to modest performance differences. Nonetheless, our independent reproduction supports the original findings that chest X-rays contain latent information about cardiac structure that a neural network can exploit. With continued refinement and external validation, such models could become useful tools for early identification of patients with occult cardiomyopathy, enabling earlier intervention and improved outcomes.

## References

Bhave, S., Rodriguez, V., Poterucha, T., Mutasa, S., Aberle, D., Capaccione, K. M., Chen, Y., Dsouza, B., Dumeer, S., Goldstein, J., Hodes, A., Leb, J., Lungren, M., Miller, M., Monoky, D., Navot, B., Wattamwar, K., Wattamwar, A., Clerkin, K., ... Elias, P. (2024). **Deep learning to detect left ventricular structural abnormalities in chest X-rays.** *European Heart Journal, 45*(22), 2002–2012. <https://doi.org/10.1093/eurheartj/ehad782>