ESC European Society of Cardiology

European Heart Journal - Digital Health (2025) **6**, 1246–1256
https://doi.org/10.1093/ehjdh/ztaf120

ORIGINAL ARTICLE

Echocardiographic measures read by artificial intelligence enable accurate and rapid prediction of the worsening of heart failure

Tony Hauptmann [1,*†], Sven-Oliver Tröbs[2,3,†], Andreas Schulz[2], Aida Romano Martinez[2,3,4], Philipp Lurz[3,5], Jürgen Prochaska[2,3,6], Philipp Sebastian Wild[2,3,4,6,†], and Stefan Kramer[1,†]

[1]Institute of Computer Science, Johannes Gutenberg University Mainz, Mainz, Germany; [2]Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany; [3]DZHK (German Center for Cardiovascular Research), Partner Site Rhine-Main, Mainz, Germany; [4]Systems Medicine, Institute for Molecular Biology (IMB), Mainz, Germany; [5]Cardiology I, Department of Cardiology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany; and [6]Clinical Epidemiology and Systems Medicine, Center for Thrombosis and Hemostasis, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

Received 1 April 2025; revised 1 July 2025; accepted 9 September 2025; online publish-ahead-of-print 15 October 2025

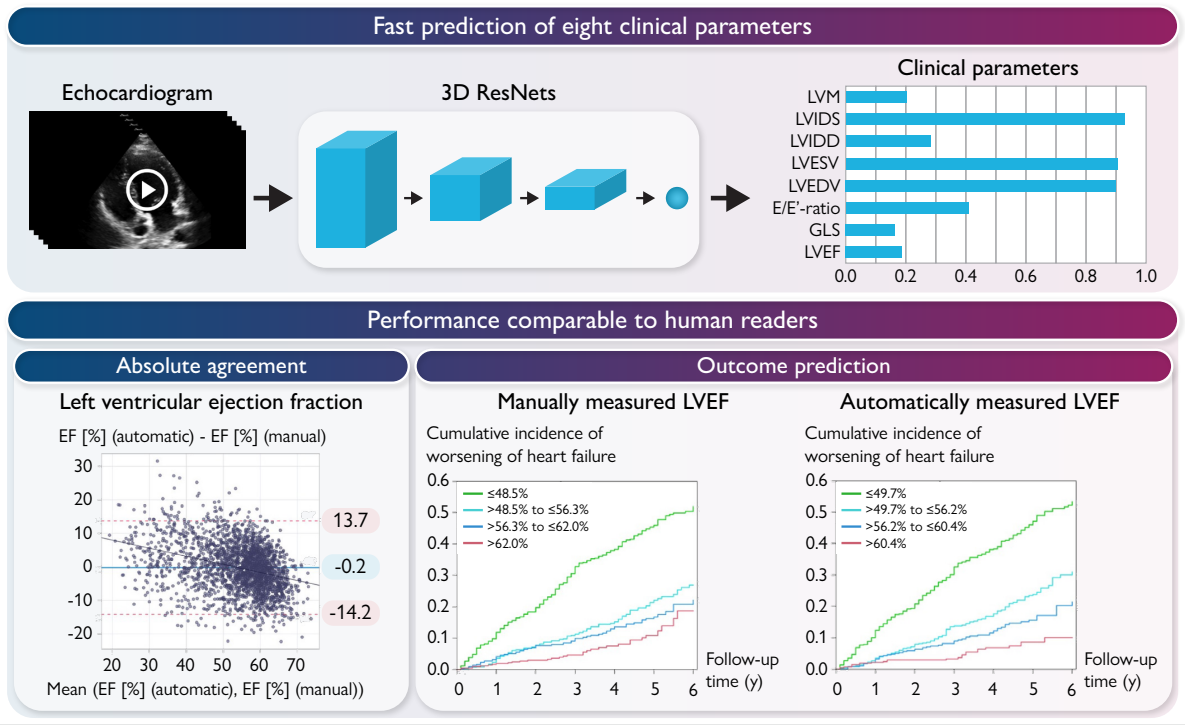
Aims Automatic echocardiographic measurements using artificial intelligence have shown promising results; however, they have not been compared with manual measurements regarding heart failure (HF) progression and algorithm runtime.

Methods and results Data came from the prospective HF study MyoVasc (NCT04064450), which involved a highly standardized 5-h examination, including comprehensive echocardiography, at a dedicated study centre between January 2013 and April 2018. Worsening of HF was a primary composite endpoint, recorded by structured follow-up, death certificates, and medical records. The automated assessment was performed using EchoDL, eight 3D convolutional neural networks (CNNs) trained to predict clinical parameters. Manual and automatic left ventricular ejection fraction (LVEF), *E*/*E*′-ratio and left ventricular mass (LVM) demonstrated a good intraclass correlation coefficient {LVEF: 0.75 [95% confidence interval (CI) 0.75–0.77], *E*/*E*′-ratio: 0.59 [CI 0.56–0.61], LVM: 0.64 [CI 0.62–0.66]}. After a median follow-up of 3.8 years (IQR 2.1–5.0), 470 patients experienced worsening of HF. In multivariable Cox analysis, comparison of manually and automatically assessed LVEF, *E*/*E*′-ratio and LVM demonstrated risk estimates slightly in favour of the CNNs. Direct comparison of *C*-indices showed significantly better model performance for automatically determined LVEF (0.71 vs. 0.73, $P = 0.038$) and *E*/*E*′-ratio (0.64 vs. 0.66, $P = 0.013$) and a trend for LVM (0.66 vs. 0.68, $P = 0.063$). Echo-DL required an average of 1053.4 ms (95% CI 1050.7–1056.0) to analyse a four-second-long echocardiogram.

Conclusion Automated analysis of echocardiograms using 3D CNNs was comparable to manual measurements in predicting HF-specific outcomes. Echo-DL offers potential time savings and improved risk prediction in clinical settings, allowing integration into echocardiographic hardware.

* Corresponding author. Tel: +49 6131 39 26901, Email: thauptmann@uni-mainz.de
†Contributed equally.

Graphical Abstract



Keywords Artificial intelligence • Echocardiography • Worsening of heart failure • Machine learning

Introduction

Echocardiography is one of the most important and frequently performed cardiological diagnostic methods, and it is essential for the diagnosis and risk stratification of heart failure (HF).[1] However, image acquisition and standard measures are time-consuming, require extensive training, and have high inter-observer variability, especially in the case of sequential examinations.[2]

Manual echocardiographic assessment relies on human visual interpretation, limiting the inclusion of information beyond the human eye's capabilities. Advanced techniques, such as global longitudinal strain (GLS) based on speckle tracking, have proven superior to conventional echocardiographic measures in the risk assessment of individuals with HF by detecting speckles inaccessible to the human eye.[3]

With the increasing computing power over the last years, the implementation of artificial intelligence has become feasible, offering the potential to improve echocardiography. Various convolutional neural networks (CNNs) achieved reliable results in view classification[4–7] and automated reading of echocardiographic measures such as left ventricular ejection fraction (LVEF) and left ventricular mass (LVM),[6] and detect conditions, e.g. atrial fibrillation. Moreover, these experiments are not limited to echocardiograms. They have also been used to detect reduced LVEF from 12-lead electrocardiograms in sinus rhythm.[8,9]

Early neural networks for echocardiographic analysis relied on segmentation data as ground truth and used 2D CNNs for frame-by-frame segmentation of each video. Based on all segmentations, an algorithm identified left ventricular end-systolic (LVESV) and end-diastolic (LVEDV) volumes, and LVEF was calculated, a time and resource-intensive approach.[6]

A recent method used automatically computed segmentations to find all frames of a heartbeat, and the corresponding video served as input for a three-dimensional (3D) CNN to predict the LVEF.[10] This decreases computational time by directly processing sequential image data, which is essential for routine clinical use.

Lau *et al.* trained a 3D CNN that embeds short segments of the echocardiogram by training a multi-task neural network for view classification and four clinical measurements. The latent representations are combined with a multi-instance attention head network to predict measurements for the study. The model accurately predicted the clinical parameters and demonstrated, in comparison to manual measures, a greater association with four incident outcomes. It showed similar quality results on two external datasets.[11] This approach is adequate for offline analysis by combining information from different videos and views, but it is less suitable for real-time echocardiographic processing.

EchoDL was developed for real-time application on standard echocardiographic hardware, enabling fast and accurate measurements during examinations. Previous work demonstrated the capabilities of neural networks to predict clinical parameters from echocardiograms, specifically for individual or a smaller number of parameters. In this study, eight functional and structural parameters were inferred from neural networks, and their correlation with clinical outcomes was analyzed, thereby increasing trust in automatic measurements.

This work focuses on developing a high-speed 3D CNN for real-time assessment on standard echocardiogram hardware. The latter part of the study evaluates the approach by comparing the predictive performance of automated measurements with manual methods, particularly

Table 1 Echocardiographic characteristics of the study sample and inter-observer reliability

Parameter	Manual	Automatic	Intraclass correlation coefficient (95% CI)
Left ventricular function			
LV ejection fraction (%)	54.2 ± 11.3	54.0 ± 9.1	0.76 (0.75–0.77)
Global longitudinal strain (%)	−17.2 ± 4.4	−17.1 ± 3.2	0.67 (0.65–0.67)
E/*E*′-ratio	9.8 ± 5.2	9.5 ± 3.6	0.59 (0.56–0.61)
Left ventricular structure			
LV end-diastolic volume (mL)	110 ± 47	109 ± 41	0.82 (0.81–0.83)
LV end-systolic volume (mL)	54 ± 37	53 ± 32	0.85 (0.84–0.86)
LV internal diameter end-diastole (cm)	5.0 ± 0.8	5.0 ± 0.6	0.66 (0.64–0.68)
LV internal diameter end-systole (cm)	3.6 ± 0.9	3.6 ± 0.7	0.77 (0.75–0.78)
LV mass (g/m)	205.0 ± 72.6	207.0 ± 56.0	0.64 (0.62–0.66)

Automatic measures were inferred with Echo-DL on the Mainz dataset with a six-fold cross-validation.
LV, left ventricular.

for HF progression. The aims of the present study were (i) to develop high-speed 3D CNNs for the assessment of echocardiograms usable on current echocardiogram hardware in real-time, (ii) to train a model that achieves equivalent performance to manual observers and to previously published neural networks, and (iii) to compare the value of manual and automated measures for predicting worsening of HF.

Methods

Study design and population

The analyzed data were from the MyoVasc study, an investigator-initiated prospective cohort study on HF (NCT04064450). Study design and baseline clinical characteristics of the study sample have been published elsewhere.[12]

In summary, a highly standardized investigation examined patients aged 35–84 with echocardiographic evidence of systolic or diastolic cardiac dysfunction, along with population-based controls, at a dedicated study centre. Of 3289 individuals categorized as HF in Stages 0 to D, according to current guidelines,[1] echocardiographic data were available in 2466 individuals. After a median follow-up of 3.8 years (IQR 2.1–5.0), the primary endpoint of worsening of HF was reached by 470 individuals, and all-cause death by 335. More information on the study sample's clinical characteristics is in Supplementary material online, *Table S1*, and the available measures for each parameter are in *Table S5*.

The responsible ethics committee and the data safety commissioner approved the study. It adhered to the Declaration of Helsinki and the standards of Good Clinical and Epidemiological Practice. The analysis of automatically derived measures was not pre-specified. Its objective was defined before implementing the experiments, which were conducted following best practices for machine learning research to mitigate bias.

Acquisition of clinical data and outcome

All study procedures were performed by trained personnel according to predefined standard operating procedures. Cardiovascular risk factors, including diabetes mellitus, dyslipidaemia, a family history of myocardial infarction or stroke, arterial hypertension, obesity, and smoking, were assessed through physical examination, anthropometry, computer-assisted interviews, blood pressure measurements, and laboratory analysis. Information on clinical outcomes was obtained through annual follow-up investigations, including a computer-assisted telephone interview, medical reports, and quarterly vital status queries to national registries. Source data, including death certificates and medical records, were obtained where possible and adjudicated by a clinical events committee. Worsening of HF was defined as the primary endpoint and included the transition from asymptomatic to symptomatic HF in asymptomatic individuals, HF hospitalization and cardiac death.

Echocardiography

Certified physicians recorded comprehensive transthoracic echocardiograms (iE33, Philips Healthcare, Germany), and images were digitally stored in an archiving and communication system (Xcelera, Philips Healthcare, Germany). All structural and functional cardiac measurements were performed according to ASE/ESC recommendations.[13] LVM was calculated using the cube formula. LVEF was determined by the Simpson method in the apical four-chamber view. Peak early (E) and late (A) diastolic inflow velocities were measured by pulsed-wave Doppler, and peak mitral annular longitudinal early velocity (*E*′) was obtained by placing the sample volume at the lateral mitral annulus. GLS was measured offline using QLab 9.0.1 (Philips Healthcare, Germany), as previously described.[3]

Data preparation and training of the 3D convolutional neural network

Available echocardiograms stored in a picture archiving system were transferred to an external storage device. DICOM metadata were removed, and each loop was converted to audio video interleave (AVI) format. The quality of apical 4-chamber (A4C) views was then categorized according to the following grades (1: Very good, 2: good, 3: satisfactory, 4: adequate, 5: poor, 6: insufficient) and labelled by a board-certified cardiologist (S.O.T.). Finally, the best A4C for obtaining left ventricular measures in a study was marked. The labelled dataset is referenced as the Mainz dataset.

The dataset was transferred to a server containing an NVIDIA A100 for training. The runtime comparison was performed on a mid-range graphics card (NVIDIA GeForce RTX 2070 SUPER) representing current echocardiography hardware.

In pre-processing, DICOM files were converted to AVI, and the interface around the heart was cropped to avoid information leakage. During the training, online augmentation was performed, including changes in brightness (reduction or increase between 0% and 10%), rotation (left or right by up to 10°), and translation (10 pixels in both horizontal and vertical directions). Additionally, temporal augmentation, where different segments of an echocardiogram were extracted and used as training input, was applied.

Eight different single-task 3D CNNs were trained, one for each clinically relevant echocardiographic measure of the Mainz dataset (LVEDV, LVESV, EF, LVM, LVIDD, LVIDS, GLS, *E*/*E*′). Two models were trained to predict multiple echocardiographic measures simultaneously: one for structural and one for functional parameters. These models will be referred to as EchoDL and multi-task EchoDL.

Table 2 3D convolutional neural network test metrics for various echocardiographic measures

Parameter	RMSE	MAE	RAE	$r_{Spearman}$
Left ventricular function				
LV ejection fraction (%)	7.104 ± 0.303	5.593 ± 0.222	0.683 ± 0.024	0.695 ± 0.027
Global longitudinal strain (%)	3.100 ± 0.161	2.449 ± 0.127	0.717 ± 0.053	0.638 ± 0.047
E/*E*′-ratio	4.066 ± 0.301	2.781 ± 0.163	0.765 ± 0.024	0.558 ± 0.040
Left ventricular structure				
LV end-diastolic volume (mL)	26.040 ± 1.033	18.970 ± 0.642	0.558 ± 0.023	0.789 ± 0.020
LV end-systolic volume (mL)	19.037 ± 1.401	12.570 ± 0.770	0.501 ± 0.018	0.802 ± 0.010
LV internal diameter end-diastole (cm)	0.586 ± 0.029	0.460 ± 0.031	0.746 ± 0.041	0.627 ± 0.032
LV internal diameter end-systole (cm)	0.554 ± 0.023	0.420 ± 0.015	0.627 ± 0.029	0.703 ± 0.026
LV mass (g/m)	54.743 ± 3.008	41.319 ± 1.595	0.743 ± 0.037	0.646 ± 0.030

The mean and standard deviation of the test sets are given. Automatic measures were inferred with Echo-DL on the Mainz dataset with a six-fold cross-validation.
LV, left ventricular; MAE, mean absolute error; RAE, relative absolute error; RMSE, root mean squared error.

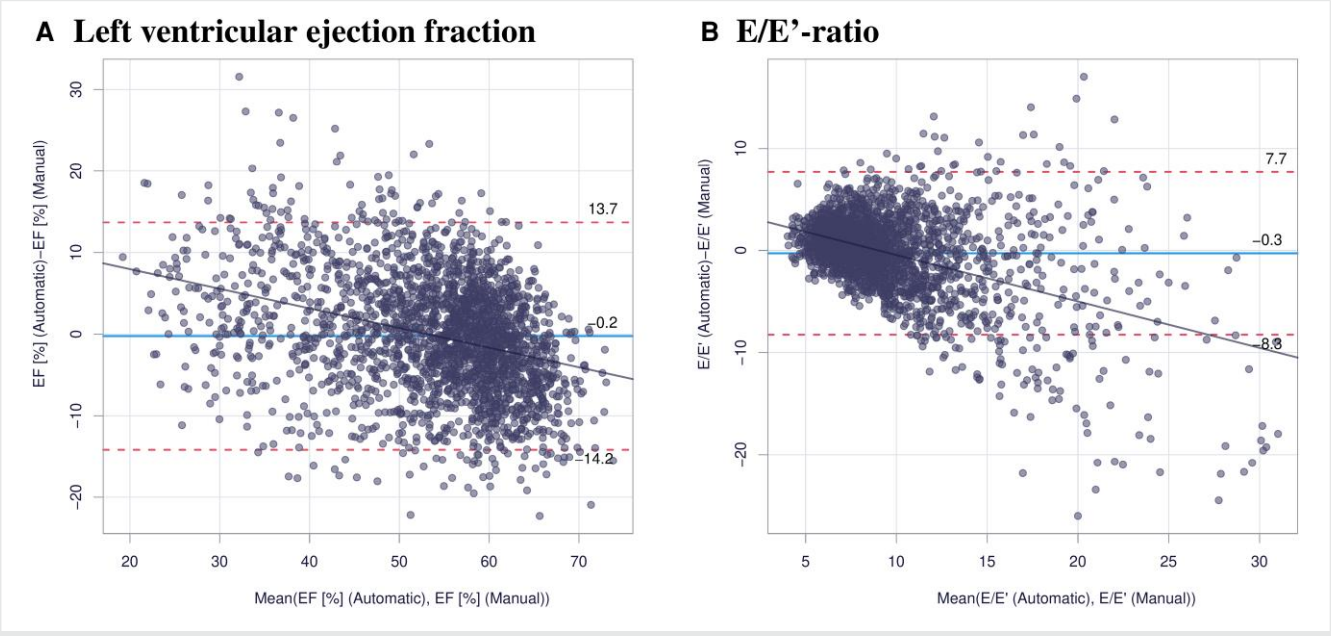


Figure 1 Comparison of automatic and manual readings by Bland–Altman plots. Bland–Altman plots of manual and automatic assessed (*A*) left ventricular ejection fraction; (*B*) *E*/*E*′-ratio.

The neural network's backbone is a 3D ResNet with 100 input frames. For inference, 100-frame subvideos were extracted from the whole echocardiogram and used as input, with the final prediction calculated as the average of all video results. Training consisted of a six-fold cross-validation on the Mainz dataset, where four sets were used for training, one for validation, and one for calculating the test metrics in each iteration.

To assess the models' performance, the mean absolute error (MAE), the root mean squared error (RMSE), and the relative absolute error (RAE) were calculated. RAE uses the mean of the training set. For external validation, the publicly accessible Stanford dataset was used. It contains 10 030 labelled videos from an A4C view, with a resolution of 112 × 112. For each echocardiogram, human expert annotations, including measurements of LVEF, LVEDV, and LVESV, as well as tracings of the left ventricle, are provided.[14]

Statistical analysis

Continuous variables are represented by mean and standard deviation for normal distributions and by median and interquartile range for skewed distributions. Relative and absolute frequencies describe discrete variables. Bland–Altman plots were generated, and intraclass correlation coefficients (ICCs) were assessed to evaluate inter-observer variability. Multivariable regression analysis was performed to analyse the association between available echocardiographic measures (manual and automated), adjusted for sex, age, and body height, with log(NT-proBNP). Cumulative incidence plots were generated, and multivariable Cox analyses, including *C*-index, were calculated for both manual and automatic echocardiographic measures, adjusted for sex and age, using worsening of HF and all-cause death as the dependent variables. Finally, the *C*-indices of automatic and manual models for the same echocardiogram were tested for differences.[15]

Table 3 Association of manually and automatically determined echocardiographic measures with log(NT-proBNP)

Parameter	Manual			Automatic			AIC (automatic) − AIC (manual)
	β-Estimate	P-value	R^2	β-Estimate	P-value	R^2	
Left ventricular function							
LV ejection fraction (%)	−0.71 (−0.75 to −0.66)	<0.0001	0.435	−0.71 (−0.75 to −0.67)	<0.0001	0.439	−14.77
Global longitudinal strain (%)	0.57 (0.53–0.62)	<0.0001	0.371	0.60 (0.56–0.65)	<0.0001	0.393	−64.53
E/*E*′-ratio	0.48 (0.44–0.53)	<0.0001	0.301	0.52 (0.47–0.56)	<0.0001	0.319	−64.32
Left ventricular structure							
LV end-diastolic volume (mL)	0.65 (0.60–0.70)	<0.0001	0.372	0.68 (0.64–0.73)	<0.0001	0.394	−82.957
LV end-systolic volume (mL)	0.71 (0.67–0.76)	<0.0001	0.427	0.70 (0.65–0.74)	<0.0001	0.421	24.16
LV internal diameter end-diastole (cm)	0.59 (0.54–0.63)	<0.0001	0.352	0.65 (0.60–0.70)	<0.0001	0.384	−123.77
LV internal diameter end-systole (cm)	0.72 (0.67–0.76)	<0.0001	0.435	0.71 (0.67–0.76)	<0.0001	0.431	17.23
LV mass (g/m)	0.59 (0.54–0.63)	<0.0001	0.342	0.66 (0.61–0.71)	<0.0001	0.380	−144.67

Multivariate linear regression analysis was adjusted for age, sex, and body height with log(NT-proBNP) as the dependent variable.
AIC, Akaike information criterion; LV, left ventricular.

Method comparison

Echo-DL was compared with two previously published methods: EchoCV[6] and EchoNet-Dynamic.[10] EchoCV employs a 2D-CNN to segment different parts of the heart and computes parameters based on the segmentation using conventional methods. EchoNet-Dynamic advances by employing a 3D CNN for parameter prediction. First, it applies a neural network to segment the left ventricle individually in each frame. The segmentations enable it to extract segments containing individual cardiac cycles by detecting frames with the largest and smallest LV area, which represent the start and end of a heartbeat. Frames from the start to the end of the heartbeat are extracted and used as input for the neural network.

This experiment aims to compare the prediction quality of the methods regardless of their training data; i.e. the comparison focuses on different approaches rather than the data used. However, training EchoNet-Dynamic on the Mainz dataset is infeasible because the dataset does not contain tracings of the left heart chamber. Therefore, the publicly available Stanford dataset, released with EchoNet-Dynamic, was used to compare the methods.

EchoNet-Dynamic and EchoDL were trained on the Stanford train split to predict EF. Since EchoCV required DICOM files as input, it could not be trained using either the Stanford or Mainz dataset. However, the published model of EchoCV, which was trained on a private dataset from the University of California, San Francisco, can be employed, albeit with the caveat that the training data differs again.

Again, the MAE, MSE, and RAE were calculated using the Stanford test split and the complete Mainz dataset to compare the models' performance. Additionally, to assess mean runtimes, 1000 repeated predictions were conducted on a four-second video comprising 200 frames.

Results

Model performance

Manual and automated echocardiographic measurements were comparable, showing only minor differences (see *Table 1*). RAE was lowest for LVEDV and LVESV, while the RAE of LVEF was slightly higher and comparable to LVM. The *E*/*E*′-ratio demonstrated a comparable RAE to LVM, even though no Doppler data were included (*Table 2*).

The Echo-DL models were compared to the multi-task Echo-DL. When predicting functional and structural measures simultaneously with two distinct models, RAE changed only marginally for most parameters. As expected, the mean runtime was reduced by ∼65% (e.g. 1053.4 s [CI 1050.7–1056.0] vs. 365.4 [CI 364.4–366.4], Supplementary material online, *Table S4*) for individual and multi-task models using the same input resolution and neural network depth.

Agreement between manual and automatic measures

The automatically obtained measures showed a moderate to strong correlation with manually read values ($r_{spearman}$ 0.558–0.802), as shown in *Table 2*.

Bland–Altman plots were created to evaluate the agreement between manual and automatic measures. The mean difference between the two methods for LVEF was −0.2 [95% line of agreement (LoA) −14.2, 13.7] and for *E*/*E*′ −0.3 (95% LoA −8.3, 7.7) (see *Figure 1*), indicating a low systematic bias. For both parameters, the slope of the linear regression is negative, particularly for *E*/*E*′. Supplementary material online, *Figures S1* and *S2* illustrate Bland–Altman plots for the remaining echocardiographic measures.

For further analysis, the ICC comparing manually and automatically read values were calculated: LVEF [ICC: 0.76 (95% CI 0.75–0.77)], end-diastolic (LVEDV, ICC: 0.82 [95% CI 0.81–0.83], and end-systolic volumes (LVESV, ICC: 0.85 [95% CI 0.84–0.86] demonstrated excellent agreement, while *E*/*E*′-ratio (ICC 0.59 [95% CI 0.56–0.61] and LVM (ICC 0.64 [95% CI 0.62–0.66] showed good agreement (*Table 1*).

Comparison of the association of manually and automatically read measures with NT-proBNP

The next step is to analyse the association of echocardiographic measures with the gold-standard HF biomarker NT-proBNP. In multivariate linear regression with log(NT-proBNP) as the dependent variable, automated echocardiographic measures, except LVESV, had slightly higher overall β-estimates than manual measurements after adjustment for age, sex, and body height (see *Table 3*).

While there was little difference for LVEF (β_{manual}: −0.71 (95% CI −0.75 to −0.66), $P < 0.00000001$ vs. $\beta_{automatic}$: −0.71 (CI −0.75 to

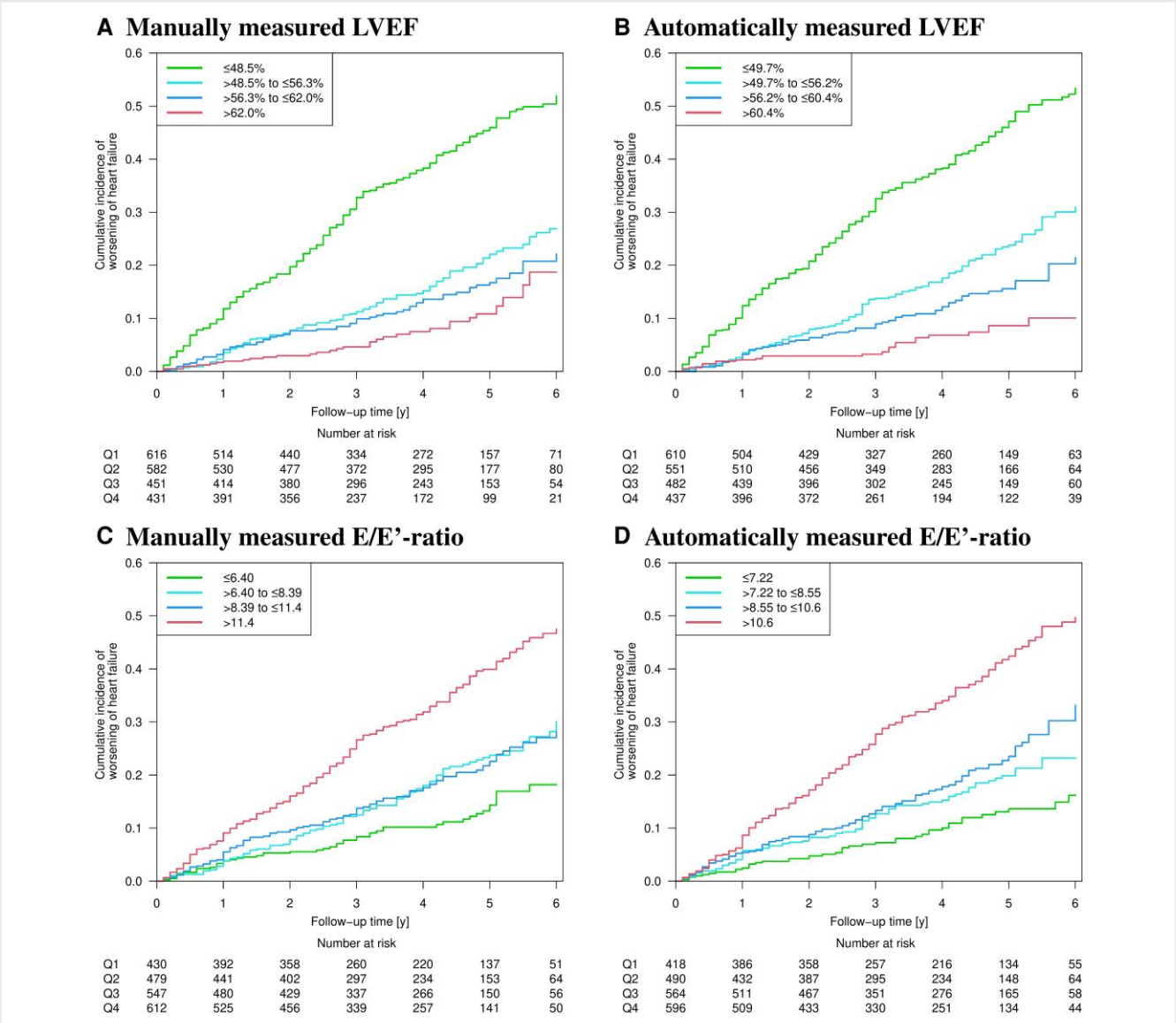


Figure 2 Incidence worsening of heart failure according to manually and automatically assessed left ventricular ejection fraction and *E*/*E*′-ratio quartiles. Cumulative incidence plots for worsening of heart failure according to quartiles of (*A*) manual, (*B*) automatic assessed left ventricular ejection fraction, (*C*) manual, and (*D*) automatic assessed *E*/*E*′-ratio.

−0.67), $P < 0.00000001$) and LVESV (β_{manual}: 0.71 (CI 0.67–0.76), $P < 0.00000001$ vs. $\beta_{automatic}$: 0.70 (CI 0.66–0.74), $P < 0.00000001$), the difference was pronounced for *E*/*E*′-ratio (β_{manual}: 0.48 (CI 0.43–0.53), $P < 0.00000001$ vs. $\beta_{automatic}$: 0.52 (CI 0.47–0.56), $P < 0.00000001$); LVM (β_{manual}: 0.59 (CI 0.54–0.64), $P < 0.00000001$ vs. $\beta_{automatic}$: 0.66 (CI 0.61–0.71), $P < 0.00000001$) and LVEDV (β_{manual}: 0.71 (CI 0.67–0.76), $P < 0.00000001$ vs. $\beta_{automatic}$: 0.70 (CI 0.66–0.74), $P < 0.00000001$).

These observations were also reflected in a higher R^2 of models that included automatically measured echocardiographic measures, except for those predicting LVESV. The *Akaike information criterion* (AIC) was used to compare the relative quality of both linear regression models. The AIC of the automatic-based regression was subtracted from that of the manual-based regression. A negative difference means the logistic regression trained on automatic measures has less information loss.

The automatic-based model was more accurate for LVEF (−14.77), GLS (−64.53), *E*/*E*′-ratio (−64.32), LVESV (−24.16), LVIDD (−123.77), and LV mass (−144.67), but the manual-based model achieved more accurate results for LVIDS (17.23) and LVEDV (82.96).

Comparison of automated and manual echocardiographic measures for outcome prediction

The applicability of manual and automatic echocardiogram assessment for predicting worsening of HF was compared. *Figure 2* shows the cumulative incidence of worsening of HF according to quartiles of (i) manually derived LVEF and (ii) automatically assessed LVEF. The curves resemble each other considerably, although the automated LVEF discriminates

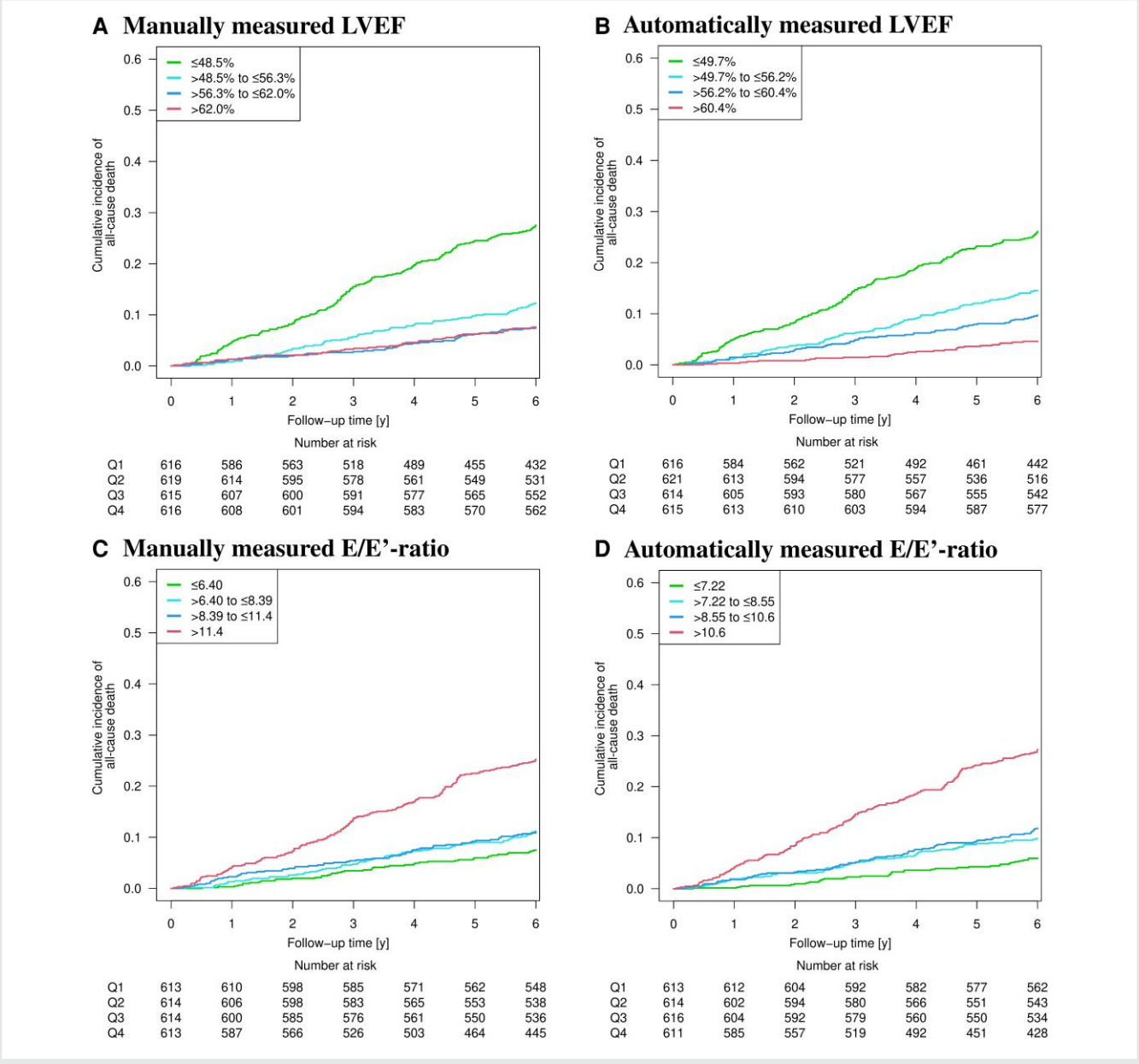


Figure 3 Incidence all-cause death according to manually and automatically assessed left ventricular ejection fraction and *E/E'*-ratio quartiles. Cumulative incidence plots for all-cause death according to quartiles of (*A*) manual, (*B*) automatic assessed left ventricular ejection fraction, (*C*) manual, and (*D*) automatic assessed *E/E'*-ratio.

slightly better within the lower three quartiles. Comparable results were observed for the *E/E'*-ratio, where the automated *E/E'*-ratio also discriminates slightly better (*Figure 2C and D*). Additional cumulative incidence plots for the remaining functional and structural echocardiographic measurements are in Supplementary material online, *Figures S3–SS5*.

For all-cause death, the cumulative incidence plots for LVEF and *E/E'*-ratio are shown in *Figure 3*. Once again, manual and automatic curves resemble each other; however, in this instance, for LVEF, the two lower quartiles are more distinct, whereas this is not the case for the *E/E'*-ratio. The remaining cumulative incidence plots are presented in Supplementary material online, *Figures S6-S8*.

To evaluate the predictive value of automatic echocardiographic measures, multivariable Cox models were constructed: After adjustment for age and sex, manually measured LVEF and *E/E'*-ratio for worsening of HF conferred a 1.84-fold [95% CI (1.71–1.99), $P < 0.0001$] and 1.31-fold [(CI 1.23–1.40), $P < 0.0001$] increased risk of worsening of HF, respectively. The risk estimates for $LVEF_{automatic}$ [hazard ratio (HR) 1.91 (CI 1.77–2.07), $P < 0.0001$] and *E/E'*-$ratio_{automatic}$ [HR 1.41 (CI 1.31–1.51), $P < 0.0001$] were higher.

Similar observations were made for cardiac structural measures: $LVEDV_{manual}$, $LVESV_{manual}$, and LVM_{manual} showed a 1.54-, 1.52-, and 1.58-fold increased risk for worsening of HF, whereas automatically obtained LVEDV, LVESV, and LVM demonstrated a 1.67-, 1.67-, and 1.75-fold increased risk, respectively, as shown in *Table 4*.

The same experiments were repeated for all-cause death (*Table 5*). For all-cause death, the differences were smaller. The indices for

Table 4 Comparison of manual and automatically determined echocardiographic measures in predicting worsening of heart failure

Parameter	Manual			Automatic			*P* for difference
	Hazard ratio [95% CI]	*P*-value	*C*-index	Hazard ratio [95% CI]	*P*-value	*C*-index	
Left ventricular function							
LV ejection fraction (%)	1.84 (1.71–1.99)	<0.0001	0.71	1.91 (1.77–2.07)	<0.0001	0.73	0.038
Global longitudinal strain (%)	2.04 (1.83–2.27)	<0.0001	0.71	1.87 (1.73–2.02)	<0.0001	0.73	0.25
E/*E*′-ratio	1.31 (1.23–1.40)	<0.0001	0.64	1.41 (1.31–1.51)	<0.0001	0.66	0.013
Left ventricular structure							
LV end-diastolic volume (mL)	1.54 (1.43–1.66)	<0.0001	0.67	1.67 (1.54–1.81)	<0.0001	0.70	0.00012
LV end-systolic volume (mL)	1.52 (1.42–1.62)	<0.0001	0.70	1.67 (1.56–1.78)	<0.0001	0.71	0.012
LV internal diameter end-diastole (cm)	1.70 (1.56–1.86)	<0.0001	0.67	1.82 (1.68–1.97)	<0.0001	0.70	0.0016
LV internal diameter end-systole (cm)	1.76 (1.63–1.88)	<0.0001	0.69	1.90 (1.77–2.05)	<0.0001	0.73	0.00014
LV mass (g/m)	1.58 (1.46–1.71)	<0.0001	0.66	1.75 (1.61–1.90)	<0.0001	0.68	0.063

Multivariable Cox regression analysis was adjusted for age and sex, with worsening of heart failure as the dependent variable.
LV, left ventricular.

Table 5 Comparison of manual and automatically determined echocardiographic measures in predicting all-cause death

Parameter	Manual			Automatic			*P* for difference
	Hazard ratio [95% CI]	*P*-value	*C*-index	Hazard ratio [95% CI]	*P*-value	*C*-index	
Left ventricular function							
LV ejection fraction (%)	0.56 (0.51–0.61)	<0.0001	0.75	0.56 (0.51–0.61)	<0.0001	0.76	0.51
Global longitudinal strain (%)	1.89 (1.67–2.15)	<0.0001	0.74	1.81 (1.65–1.99)	<0.0001	0.75	0.78
E/*E*′-ratio	1.45 (1.35–1.56)	<0.0001	0.73	1.46 (1.36–1.58)	<0.0001	0.74	0.25
Left ventricular structure							
LV end-diastolic volume (mL)	1.51 (1.38–1.64)	<0.0001	0.73	1.58 (1.44–1.72)	<0.0001	0.73	0.30
LV end-systolic volume (mL)	1.53 (1.43–1.65)	<0.0001	0.74	1.55 (1.44–1.68)	<0.0001	0.74	0.43
LV internal diameter end-diastole (cm)	1.50 (1.36–1.67)	<0.0001	0.72	1.65 (1.50–1.82)	<0.0001	0.74	0.061
LV internal diameter end-systole (cm)	1.69 (1.55–1.85)	<0.0001	0.74	1.69 (1.54–1.85)	<0.0001	0.74	0.97
LV mass (g/m)	1.52 (1.38–1.67)	<0.0001	0.73	1.66 (1.51–1.83)	<0.0001	0.74	0.26

Multivariable Cox regression analysis was adjusted for age and sex, with worsening of heart failure as the dependent variable.
LV, left ventricular.

automatic values are greater in five cases and the same in the remaining cases; however, there is no evidence of statistical differences.

Model performance was compared using the *C*-index to measure differences in HRs. Except for GLS, the *C*-indices of the models based on automated measures were higher. Indeed, a statistically significant difference could be confirmed for the functional parameters LVEF and *E*/*E*′-ratio, as well as for the structural parameters LVEDV, LVESV, LVIDS, and LVIDD. At the same time, a trend was observed for LVM (*Table 4*).

Method comparison

First, the predictive performance of LVEF was compared across three methods (EchoDL, EchoNet-Dynamic, and EchoCV) and two different datasets (Mainz and Stanford).

For the Mainz dataset, the test metrics (MAE, RMSE, and RAE) of EchoDL were comparable to those of EchoNet-Dynamic, as shown in *Table 6*. The LVEF model, with an input resolution of 112 × 112, was evaluated on the Stanford dataset for external validation, achieving comparable results: MAE 5.908, RMSE 7.906, RAE 0.562, $r_{Spearman}$ 0.678 (*Table 6*).

However, EchoCV demonstrated error metrics nearly twice as high as those of EchoNet-Dynamic and EchoDL. Unfortunately, it is impossible to verify whether the differences stem from the training data or the method used by EchoCV.

EchoCV required the longest prediction time at 22.76 s (95% CI 22.72–22.79). EchoNet-Dynamic was the second fastest method, at 394.7 ms (CI 394.3–395.1), while EchoDL was approximately four times faster, with 102.5 ms (CI 102.0–103.0).

Discussion

This study evaluated eight 3D CNNs for the automated measurement of structural and functional left ventricular parameters. Echo-DL was benchmarked against two existing methods and compared with manually obtained measures for predicting NT-proBNP and worsening of HF.

Table 6 Comparison of convolutional neural networks for echocardiographic assessment on a single GPU

Method	Training set	Test set	MAE	RMSE	RAE	$r_{Spearman}$
EchoDL	Mainz dataset	Stanford dataset	5.908	7.906	0.562	0.678
EchoDL	Stanford training set	Mainz dataset	5.155	6.621	0.618	0.715
EchoNet-Dynamic	Stanford training set	Mainz dataset	5.311	6.788	0.637	0.726
EchoCV	Pre-trained on a private dataset	Mainz dataset	9.350	11.611	1.005	0.536
EchoDL	Stanford training set	Stanford test set	4.179	5.581	0.463	0.819
EchoNet-Dynamic	Stanford training set	Stanford test set	4.198	5.535	0.460	0.811

EchoDL and EchoNet-Dynamic were trained using the Stanford training set. EchoCV was trained on the private University of California, San Francisco dataset.[5]
GPU, graphics processing unit; MAE, mean absolute error; RAE, relative absolute error; RMSE, root mean squared error.

Echo-DL is a simplified version of EchoNet-Dynamic that does not depend on segmentation to detect individual heartbeats. Instead, it uses enough input frames to encompass an entire heartbeat, which is expected to enhance processing speed, achieving near real-time predictions. Echo-DL demonstrated error rates comparable to those of the other models while requiring the least computational time, thereby validating that the extraction of individual heartbeats is unnecessary.

EchoDL's low computational requirements enable it to run on off-the-shelf echocardiography hardware without needing a high-end GPU. To estimate the time required for manual measurements, two study clinicians measured the same parameters five times in an offline setting. Although this approach was not performed on the same echocardiograms, making a direct comparison impossible, it does enable the estimation of potential time improvements. The study physicians spent ~11.7 min (see Supplementary material online, *Material*) measuring the same parameters, reducing the required time from minutes to seconds. Plans are being made to implement the model on echocardiographic equipment in accordance with existing regulations.

The automated parameter prediction supports clinicians, rather than replacing them. It provides clinical parameters comparable to those measured manually, enhancing clinicians' efficiency and reallocating time from data acquisition to patient-centric decision-making and care.

The ICC for the automatically predicted structural and functional measures was high compared to the manually obtained measures. Remarkably, besides volume parameters (e.g. LVEDV and LVESV) routinely assessed from the A4C view, the model was able to predict measures for which other views are recommended in the current guidelines, such as LVM and GLS,[13] demonstrating good agreement with readings obtained manually from the recommended view.

Additionally, EchoDL could predict Doppler-based indices from B-scan imaging data without tissue Doppler or pulsed-wave data. Although the temporal resolution of B-scan imaging data is low, the correlation between automatically and manually assessed *E*/*E*′-ratios was high.[16]

In the Bland–Altman analyses, automatically determined echocardiographic parameters demonstrated the best agreement with manual measurements when the values were within the normal range. The lower model performance around pathological values is undoubtedly attributable to the underlying dataset: More pathological echocardiograms for model training could enhance model accuracy[17] in these areas. Conversely, EchoDL can likely differentiate between normal and pathological readings with a slightly modified training procedure, which would greatly benefit routine clinical practice.

The negative slopes in the Bland–Altman plots suggest that automatic measurements tend to be more conservative and do not frequently predict values far from the mean. This is likely due to the small size of the training data, and more extensive training data could improve predictions in this area.

In the last experiments, the association between automatically derived measures and NT-proBNP, as well as the worsening of HF and all-cause death, was validated. Generally, the automatic measures showed a stronger association with NT-proBNP, as reflected by higher effect estimates and model performance.

Automatically predicted GLS achieved a higher *C*-index, but it was associated with a lower HR. Except for GLS, all automatic measures indicate a higher risk for worsening of HF compared to manual measures, as shown by higher HRs and *C*-indices. Despite statistically significant differences in *C*-indices in most cases, these differences are not distinctive enough to enhance the prediction of worsening of HF in general. However, neural networks have the advantage of providing predictions comparable to those of expert readers while reducing variability and saving time in clinical practice. Additionally, there are situations where neural networks enable the measurement of clinical parameters from echocardiograms that would be impossible to obtain otherwise, such as when sonographers are available but expert readers are not.

In summary, automatically derived measures correlate with worsening of HF, NT-proBNP, and manual measurements, while providing faster processing times. The neural networks achieved quality comparable to that of manual raters. Additionally, they reduced the occasional inaccuracies inherent in manual measurements, leading to improvements, particularly in scenarios involving extensive assessments. Although there is concern regarding machine learning models making erroneous predictions, this study can improve confidence in their use for echocardiogram assessment.

Limitations

EchoDL was trained using a highly standardized and in-depth phenotyped cohort of patients with symptomatic and asymptomatic HF. Trained physicians performed echocardiograms using predefined standard operating procedures, and the echocardiograms were digitally stored with minimal compression. MyoVasc includes an extensive HF-specific long-term follow-up, allowing a direct comparison of automated and manual echocardiographic measurements. The prediction of LVEF has been externally validated with reliable results; however, other measures have not yet undergone external validation.

Without a specific approach, neural networks, including EchoDL, provide only a point prediction, lacking information about uncertainties. This prevents clinicians from evaluating the reliability of a prediction. Uncertainty quantification methods provide lower and upper bound estimations, thereby improving the ability to assess the model's confidence.[18]

The results were rigorously validated and compared to increase confidence; however, the neural networks do not directly provide means to interpret or explain their results. Nevertheless, the explainability of neural networks for video analysis, particularly in medicine, remains an ongoing research topic with no widely applicable methods.[19,20] The challenge is to create explanations that are quick to interpret and do not hinder performance. In the context of explainability for echocardiograms, initial research focused on individual clinical parameters, particularly LVEF. It included methods such as generating counterfactuals[21] or a 3D depth map of the left ventricle.[22]

Conclusions

The automatic assessment of echocardiograms using machine learning is comparable to that of conventional assessment and is completed in less time. Moreover, the automatically measured parameters showed a stronger correlation with the prediction of worsening of HF and NT-proBNP levels than manual measurements. Using machine learning as a decision support tool in echocardiogram assessment enhances efficiency by saving time. It can potentially improve patient outcomes through more reliable and consistent evaluations.

Lead author biography

Tony Hauptmann has been a research associate in the Data Mining Group led by Professor Kramer at the Computer Science Institute of Johannes Gutenberg University Mainz in Germany since April 2020. He received his Master of Science in Computer Science from the University of Münster. His current research focuses on the development and integration of machine learning techniques into the life sciences.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Acknowledgements

We thank all the MyoVasc Study participants for their commitment, the study centre's clinical staff, and all colleagues who contributed to the successful implementation of this project. We thank Alexander Gieswinkel for his support.

Author contributions

Tony Hauptmann (Formal Analysis, Investigation, Methodology, Software, Validation, Writing—original draft), Sven-Oliver Tröbs (Conceptualization, Data curation, Methodology, Software, Validation, Writing—original draft), Andreas Schulz (Formal Analysis), Aida Romano Martinez (Formal Analysis), Philipp Lurz (Resources), Jürgen Prochaska (Resources), Philipp Sebastian Wild (Conceptualization, Funding acquisition, Supervision, Writing—original draft), and Stefan Kramer (Conceptualization, Funding acquisition, Methodology, Supervision, Writing—original draft).

Funding

The Bundesministerium für Bildung und Forschung funded this work and research as part of the DIASyM project under grant numbers 031L0217A and 161L0217A.

Conflict of interest: P.S.W. is principal investigator of the DIASyM research core, which focuses on the study of the HF syndrome (BMBF 161L0217A) and principal investigator of the future cluster 'curATime' (BMBF 03ZU1202AA, 03ZU1202CD, 03ZU1202DB, 03ZU1202JC, 03ZU1202KB, 03ZU1202LB, 03ZU1202MB, and 03ZU1202OA). P.S.W. reports no disclosures for the submitted work. Outside the submitted work, he reports on non-financial grants from Philips Medical Systems, grants and consulting fees from Boehringer Ingelheim, grants and consulting fees from Novartis Pharma, grants and consulting fees from sanofi-aventis, grants, consulting, and lecturing fees from Bayer Health Care, grants and consulting fees from Daiichi Sankyo Europe, lecturing fees from Pfizer Pharma, lecturing fees from Bristol Myers Squibb, consulting fees from Astra Zeneca, consulting fees and non-financial support from DiaSorin and non-financial support from I.E.M.

Data availability

The data underlying this article cannot be shared publicly to protect the privacy of the individuals who participated in the study.

References

1. Bozkurt B, Coats AJS, Tsutsui H, Abdelhamid CM, Adamopoulos S, Albert N, *et al.* Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure: Endorsed by the Canadian Heart Failure Society, Heart Failure Association of India, Cardiac Society of Australia and New Zealand, and Chinese Heart Failure Association. *Eur J Heart Fail* 2021;**23**:352–380.
2. Pellikka PA, She L, Holly TA, Lin G, Varadarajan P, Pai RG, *et al.* Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. *JAMA Netw Open* 2018;**1**:e181456.
3. Tröbs SO, Prochaska JH, Schwuchow-Thonke S, Schulz A, Müller F, Heidorn MW, *et al.* Association of global longitudinal strain with clinical status and mortality in patients with chronic heart failure. *JAMA Cardiol* 2021;**6**:448–456.
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p:770–778. https://doi.org/10.1109/CVPR.2016.90
5. Østvik A, Smistad E, Espeland T, Berg EAR, Lovstakken L. Automatic myocardial strain imaging in echocardiography using deep learning. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, eds. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing; 2018. p309–316. (Lecture Notes in Computer Science, vol. 11045). Available from: https://link.springer.com/10.1007/978-3-030-00889-5_35
6. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, *et al.* Fully automated echocardiogram interpretation in clinical practice. *Circulation* 2018;**138**:1623–1635.
7. Østvik A, Smistad E, Aase SA, Haugen BO, Lovstakken L. Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. *Ultrasound Med Biol* 2019;**45**:374–384.
8. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, *et al.* An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;**394**:861–867.
9. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, *et al.* Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019;**25**:70–74.
10. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020;**580**:252–256.
11. Lau ES, Achille PD, Kopparapu K, Andrews CT, Singh P, Reeder C, *et al.* Deep learning–enabled assessment of left heart structure and function predicts cardiovascular outcomes. *J Am Coll Cardiol* 2023;**82**:1936–1948.
12. Göbel S, Prochaska JH, Tröbs SO, Panova-Noeva M, Espinola-Klein C, Michal M, *et al.* Rationale, design and baseline characteristics of the MyoVasc study: a prospective cohort study investigating development and progression of heart failure. *Eur J Prev Cardiol* 2021;**28**:1009–1018.
13. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, *et al.* Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J Am Soc Echocardiogr* 2015;**28**:1–39.e14.
14. Ouyang D, He B, Ghorbani A, Lungren MP, Ashley EA, Liang DH, *et al.* EchoNet-Dynamic: a Large New Cardiac Motion Video Data Resource for Medical Machine Learning. *NeurIPS* 2019 Workshop on Machine Learning for Health.
15. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med* 2015;**34**:685–703.
16. Cameli M, Mondillo S, Solari M, Righini FM, Andrei V, Contaldi C, *et al.* Echocardiographic assessment of left ventricular systolic function: from ejection fraction to torsion. *Heart Fail Rev* 2016;**21**:77–94.

17. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, *et al.* Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol* 2019;**73**:1317–1335.
18. Khosravi A, Nahavandi S, Creighton D, Atiya AF. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Trans Neural Netw* 2011;**22**:337–346.
19. Saha A, Gupta S, Ankireddy SK, Chahine K, Ghosh J. Exploring Explainability in Video Action Recognition. 2024. p8176–8181. https://openaccess.thecvf.com/content/CVPR2024W/XAI4CV/html/Saha_Exploring_Explainability_in_Video_Action_Recognition_CVPRW_2024_paper.html (23 May 2025).
20. Kolarik M, Sarnovsky M, Paralic J, Babic F. Explainability of deep learning models in medical video analysis: a survey. *PeerJ Comput Sci* 2023;**9**:e1253.
21. Reynaud H, Vlontzos A, Dombrowski M, Gilligan Lee C, Beqiri A, Leeson P, *et al.* D'ARTAGNAN: counterfactual video generation. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*. Cham: Springer Nature Switzerland; 2022. p599–609.
22. Duffy G, Jain I, He B, Ouyang D. Interpretable deep learning prediction of 3d assessment of cardiac function. *Pac Symp Biocomput* 2022;**27**:231–241.