Computers in Biology and Medicine 171 (2024) 108192

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Automated mitral inflow Doppler peak velocity measurement using deep learning

Jevgeni Jevsikov [a,b,*], Tiffany Ng [b], Elisabeth S. Lane [a], Eman Alajrami [a], Preshen Naidoo [a], Patricia Fernandes [a], Joban S. Sehmi [c], Maysaa Alzetani [d], Camelia D. Demetrescu [e], Neda Azarmehr [a], Nasim Dadashi Serej [a], Catherine C. Stowell [b], Matthew J. Shun-Shin [b], Darrel P. Francis [b], Massoud Zolgharni [a,b]

[a] School of Computing and Engineering, University of West London, United Kingdom
[b] National Heart and Lung Institute, Imperial College London, United Kingdom
[c] West Hertfordshire Hospitals NHS Trust, Wafford, United Kingdom
[d] Luton & Dunstable University Hospital, Bedfordshire, United Kingdom
[e] Luton & Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

ARTICLE INFO

Keywords:
Automated analysis
Deep learning
Doppler echocardiography
Mitral inflow

ABSTRACT

Doppler echocardiography is a widely utilised non-invasive imaging modality for assessing the functionality of heart valves, including the mitral valve. Manual assessments of Doppler traces by clinicians introduce variability, prompting the need for automated solutions. This study introduces an innovative deep learning model for automated detection of peak velocity measurements from mitral inflow Doppler images, independent from Electrocardiogram information. A dataset of Doppler images annotated by multiple expert cardiologists was established, serving as a robust benchmark. The model leverages heatmap regression networks, achieving 96% detection accuracy. The model discrepancy with the expert consensus falls comfortably within the range of inter- and intra-observer variability in measuring Doppler peak velocities. The dataset and models are open-source, fostering further research and clinical application.

1. Introduction

In contemporary cardiac research, the predominant method for evaluating ventricular filling is through pulsed-wave Doppler echocardiography. This modality quantifies the transmitral velocity during diastole, comprising of two distinct components: (i) the E-wave (measured at the peak early phase of ventricular filling), which is started by active mechanical suction of blood from the atrium by the recoiling and simultaneously relaxing ventricle, and (ii) the A-wave (measured at the peak of late diastolic filling corresponding to the atrial phase of ventricular filling), caused by the contraction of the left atrium, which caps off the ventricle and raises its pressure and volume.

At present, velocity assessments on Doppler traces are predominantly executed manually by clinicians, resulting in substantial intra- and inter-observer variability [1–3]. It has been demonstrated that human factors are the source of the error in peak Doppler velocity measurements [4].

Given the protracted nature of manual analysis, such manual assessments frequently focus on a singular heartbeat, rather than a combination of heartbeats with subsequent averaging, which has been recommended by the current clinical guidelines [5].

The integration of an automated systems could potentially facilitate the standardisation of these measurement methodologies, thereby reducing test–retest variability, and optimising clinical workflow. The main goal of the present study was therefore to develop a pipeline for the automated measurement of the Doppler transmitral velocity peaks, and to validate this method against the gold-standard manual measurements, procured from human specialists.

This paper begins by reviewing existing work on automated Doppler mitral inflow measurements. We then outline our key contributions in this area. The methodology section details our dataset, ground-truth definitions, and deep learning framework. In the results and discussion, we analyse model performance, observer variability, and note limitations. The paper concludes with a summary of our findings and potential future research directions.

* Corresponding author at: School of Computing and Engineering, University of West London, United Kingdom.
E-mail address: Jevgeni.Jevsikov@uwl.ac.uk (J. Jevsikov).

https://doi.org/10.1016/j.compbiomed.2024.108192
Received 29 September 2023; Received in revised form 1 February 2024; Accepted 18 February 2024
Available online 23 February 2024

1.1. Related work

Basic image processing methods — Early attempts to automate Doppler measurements in echocardiography primarily used basic signal and image processing techniques [2,6–11]. These methods encompassed traditional image processing tasks like noise filtering and edge detection to acquire the Doppler envelope, and thresholding for identifying key points essential for extracting clinical measurements.

For instance, Taebi et al. [10] presented a threshold-based method for extracting positive and negative peak velocity profiles from Doppler images. The method involves manual determination of the Doppler region of interest (ROI), calculation of average pixel intensity within the ROI, and smoothing using moving average. Using two thresholding methods, pixel edges are detected, and positive and negative peak velocity profiles are constructed by connecting upper and lower edges. Experimental results highlight the method's efficiency and computational advantages over edge detection methods like Prewitt and Canny.

Kiruthika et al. [11] introduced an image processing approach for delineating Doppler envelopes and calculating peak velocity. The method includes semi-automated Doppler region localisation, filtering, and application of the Canny edge detector for spectral envelope segmentation. The highest peak value is then detected by scanning the curve.

Additional studies in this category can be explored in the works of Biradar et al. [12], Syeda-Mahmood et al. [13], Greenspan et al. [14], and Shechner et al. [15].

Despite their foundational role, these algorithms encountered limitations related to suboptimal contrast and image artifacts. Additionally, their effectiveness was impeded by the necessity for meticulous hyperparameter tuning tailored to specific views. Consequently, these challenges hindered the development of a robust and universally applicable automated Doppler analysis.

Deep learning approaches — In recent years, there has been a significant shift towards the application of advanced Deep Learning (DL) techniques, including convolutional neural networks (CNNs), for automating mitral inflow measurements. These deep neural networks have shown promise in improving the accuracy and efficiency of this critical cardiac assessment.

Zamzmi et al. [16] used Faster R-CNN to extract Electrocardiogram (ECG) signals, which were used to segment each Doppler image into individual beats. However, the cardiac beat segmentation was done manually in about 10% of their images because the automatic segmentation failed due to large overlap between Doppler and ECG signals or noisy ECG signal with multiple peaks. They then used Machine Learning approaches such as K-means clustering algorithm and Gradient Vector Flow driven snake for spectral envelope delineation, and peak detection. The use of a small patient dataset (only 701 images which included different Doppler modalities) and the algorithm's dependence on the ECG were the main limitations of this study.

Elwazir et al. [17] adopt a different method. They train a deep learning classifier to differentiate between echocardiographic study types. In order to derive the envelope profile, mitral inflow images are segmented using a U-Net network. The beats were distinguished using ECG tracings. The E-and A-waves were then detected by signal processing of the segmented envelope. They reported mean velocity error of 0.06 ± 0.03 m/s and 0.05 ± 0.03 m/s for E-wave and A-wave, respectively. The fact that only echo images of normal patients without abnormalities were included in the training data, their system was dependent on the ECG signal to detect the heartbeats, and they used only one set of manual measurements to examine the performance of their proposed model, were a significant limitation of the study.

Jahren et al. [18] investigated methods for isolating heartbeat cycles on cardiac spectral Doppler spectrograms independent of an ECG signal. They combined a CNN module that collected local features from an image with a Recurrent Neural Network (RNN) module that connected the extracted features temporally. The heartbeat detection model attained an accuracy rate of 97.7% for accurate detections and a false detection rate of 2.5%. However, this study did not go beyond identifying the end-diastole locations in spectral Doppler spectrograms.

Yang et al. [19] developed a framework using DL-techniques to detect valvular heart diseases, as part of which, they calculated the mitral valve area. The method involves segmenting the Doppler waveform and extracting boxes containing one heartbeat from the segmented waveform. Extracted boxes are fed to HRNet [20] to obtain two keypoints. Similarly, for Aortic Stenosis, the maximum blood flow velocity was obtained by simply taking a maximum value from the segmented waveform. Despite their framework being able to perform complex analysis of valvular heart diseases, measurements involving Doppler mitral inflow involved multiple steps: separation of the Doppler waveform into boxes and only then passing these boxes into a keypoint detection model, which could predict a fixed number of keypoints.

We have previously reported on a DL-based ECG-independent technique to isolate heartbeats and localise blood flow velocity peaks on Tissue Doppler Images [21].

Summary — This literature survey highlights the progression from basic signal and image processing to advanced Deep Learning techniques in automating Doppler measurements in echocardiography. Despite these achievements, challenges persist, including reliance on ECG signals, need for manual intervention in semi-automated approaches, limited datasets, and complex multi-step processes. The survey indicates a need for further improvement in developing more integrated, efficient, and universally applicable models for Doppler measurement automation in echocardiography.

1.2. Main contributions

To the best of our knowledge, no approach utilises current state-of-the-art DL methods for fully automated and ECG-free estimation of peak velocities from mitral inflow Doppler images. This study, therefore, presents an innovative deep learning model designed for the automated detection of peak velocity measurements from mitral inflow Doppler images. The study's main contributions are as follows:

- Creation of a dataset of Doppler images, annotated by accredited and experienced echocardiography experts, which has been made publicly available through this report for use in developing automated models.
- Creation of a majority-based consensus dataset to serve as a uniquely robust and representative benchmark for performance evaluations.
- Demonstration of the feasibility of using CNNs to reliably detect and measure mitral inflow Doppler peak velocities, independent of ECG information.
- Public release of our developed codes and DL models, which not only provide a benchmark for future studies but also enable external validation of our reported models.

2. Method

2.1. Patient datasets and expert annotations

Fig. 1 presents a visual flowchart outlining the steps involved in preparing the dataset for this study.

Development-dataset — A large random sample of 1064 echocardiographic studies from different patients, conducted in years between 2010 and 2016, was extracted from the echocardiogram database of Imperial College Healthcare NHS Trust. The echocardiograms were acquired during examinations performed by experienced echocardiographers, following standard protocols. Ethical approval was obtained from the Health Regulatory Agency for the anonymised export of large

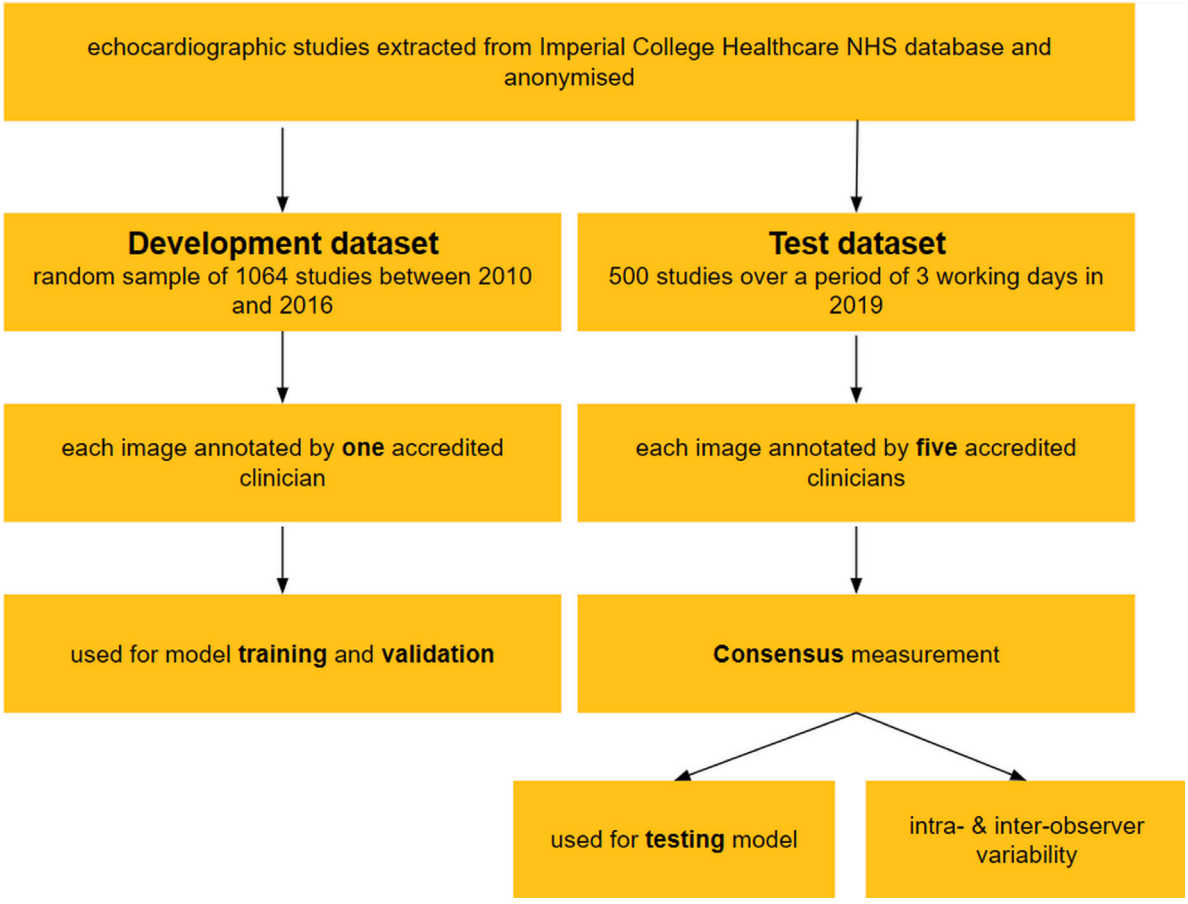

Fig. 1. Flowchart demonstrating the pipeline for preparing the two datasets used in this study; ethical approval was obtained from the Health Regulatory Agency.

quantities of imaging data. As the data was originally acquired for clinical purposes, individual patient consent was not required.

Still pulsed wave mitral inflow Doppler images were automatically extracted from each DICOM-formatted echo exam using our previously developed echo view classifier [22]. Automated anonymisation was then performed to remove the patient-identifiable information.

Next, utilising our online labelling platform (https://unityimaging.net), a sample snapshot from which is shown on Fig. 2, each image underwent labelling once. A pool of accredited and experienced clinical experts marked the E- and A-wave velocity keypoints on the Doppler images. The experts were given instructions to annotate all visible peak velocity points across several heartbeats present in each image, excluding low-quality points that they deemed unsuitable for clinical practice. This labelled dataset was utilised for model developments.

The model development dataset (images and labels) are available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license at https://intsav.github.io/doppler.html [23]. The release of associated dataset received a Favourable Opinion from the South Central – Oxford C Research Ethics Committee (Integrated Research Application System identifier 279328, 20/SC/0386).

Test-dataset — The testing dataset was curated from a series of investigations conducted over a period of three working days in 2019, years away from the development dataset. The testing set is composed of 200 Doppler images.

This research adopts a unique approach to model evaluation by leveraging a consensus testing dataset which capitalises on the collective measurements of multiple experts. For each image in the test dataset, we acquired annotations from multiple experts, resulting in a rich array of data points that encapsulates a variety of perspectives. From these multiple annotations, we derived a consensus measurement for each image, representing the majority agreement among the experts.

Five experts labelled each image using the same platform (2 of the 5 experts also re-annotated at least one heartbeat on each image, allowing the measurement of intra-observer variability). The images were presented in a random order, and each expert was blinded to any previous labelling by themselves or others. Again, the experts were instructed to label every peak unless the image quality rendered it impossible to do so. This provided us the high quality consensus reference measurements which could also be used for examining the inter- and intra-observer variability.

To generate the consensus dataset necessary for our evaluation, we employed the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [24,25] algorithm, an efficient and effective method for discovering clusters of arbitrary shape in spatial data. DBSCAN has the advantage of not requiring a predetermined number of clusters, making it particularly suitable for our application where clusters (corresponding to peak velocities) may vary across images.

Critical to our implementation of DBSCAN were two parameters: epsilon and MinPts. The epsilon parameter, which defines the maximum distance between two samples for them to be considered as in the same cluster, was set to a distance of 30 pixels. This value was chosen to reflect the reasonable expectation of proximity in experts' annotations for a given peak. The MinPts parameter, the minimum number of samples in a neighbourhood for a point to be considered as a core point, was set to the majority number of experts. This ensured that a core point, and hence a cluster, was established only when the majority of experts agreed on an annotation for a peak.

The clustering process was based solely on the x-axis coordinates of the experts' annotations, postulating that for each peak, experts

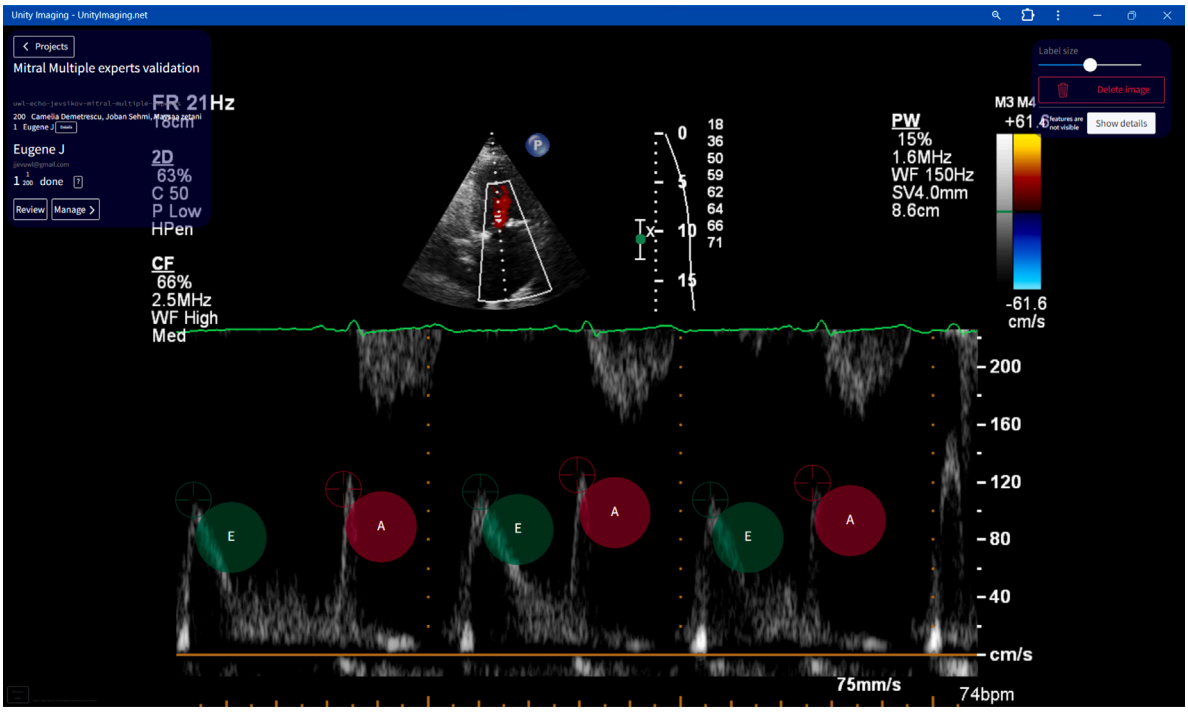

Fig. 2. A snapshot of Unity platform, used for labelling in this study. Platform is accessible at https://unityimaging.net.

should make proximate annotations to each other, and there should not be any repetitive annotations for a peak from any expert. Each peak was anticipated to be annotated once by each expert, facilitating the clustering operation.

Our use of DBSCAN focused on identifying and utilising only core points within each cluster. This approach served to exclude potential outliers or less-agreed upon points that might be included as border points, thereby enhancing the precision and robustness of the consensus.

Once clusters were identified, we calculated the consensus measurement for each cluster by averaging the coordinates of all core points within that cluster. This method yielded a consensus value for each peak velocity in an image, effectively synthesising the collective insights of the expert annotators.

This consensus-based approach offers several significant benefits. Firstly, it mitigates the potential bias associated with a single expert's interpretation, thereby augmenting the reliability and objectivity of the evaluation metrics. Secondly, by encapsulating the natural variability and nuances among expert annotations, the consensus set offers a more comprehensive and representative 'gold standard' for performance evaluation. Finally, this method enables the assessment of the model's alignment with a broader expert community rather than a single perspective, facilitating a more robust and thorough validation of the model's accuracy and reliability.

This dataset is being kept private exclusively for competition purposes.

2.2. Ground-truth definition

A multi-stage heatmap regression network is utilised to detect the peak velocities. Instead of predicting Cartesian coordinates, the model uses a different Gaussian response heatmap or belief map for every keypoint of interest. This heatmap is an image that shows the probability of a specific keypoint residing at a pixel. The model then obtains keypoints by identifying the local maximums in the heatmaps. Indirect inference through a predicted heatmap provides several advantages over direct prediction [26].

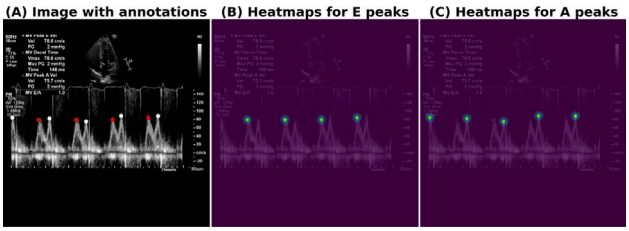

Fig. 3. Example of Ground-truth definition: (A) Original image with manually annotated peak velocities where red and white circles indicate E- and A-waves, respectively; (B, C) corresponding generated heatmaps (overlayed on the original image for visualisation purposes), used as Ground-truth.

A set of heatmaps was developed for each image that served as the Ground-Truth. For each image, there were two heatmaps: one representing the Ground-truth for E-wave, and the other for A-wave; forming a 2-channel Ground-truth for the neural network.

At each manually identified wave peak coordinate, a symmetric Gaussian distribution with a standard deviation σ was created to form the heatmaps. For each input image $I \in R^{W \times H}$,

Gaussian distributions are generated using Eq. (1).

$$Y_{ijk} = exp(-\frac{(x - G_{ik})^2 + (y - G_{jk})^2}{2\sigma^2}) \quad (1)$$

Here, σ is a size-adaptive standard deviation, Y_{ijk} is a heatmap representation of landmark coordinate, where kth is a channel index (E-wave or A-wave), and G_{ik} along with G_{jk} are the ground-truth landmark coordinates in each k channel.

Fig. 3 illustrates typical generated heatmaps with Gaussian peaks. The spread of the Gaussian peaks, which is controlled by the standard deviation σ value, was set to 5, which is sufficient to precisely narrow down and localise landmarks while avoiding any irrelevant background information from the image.

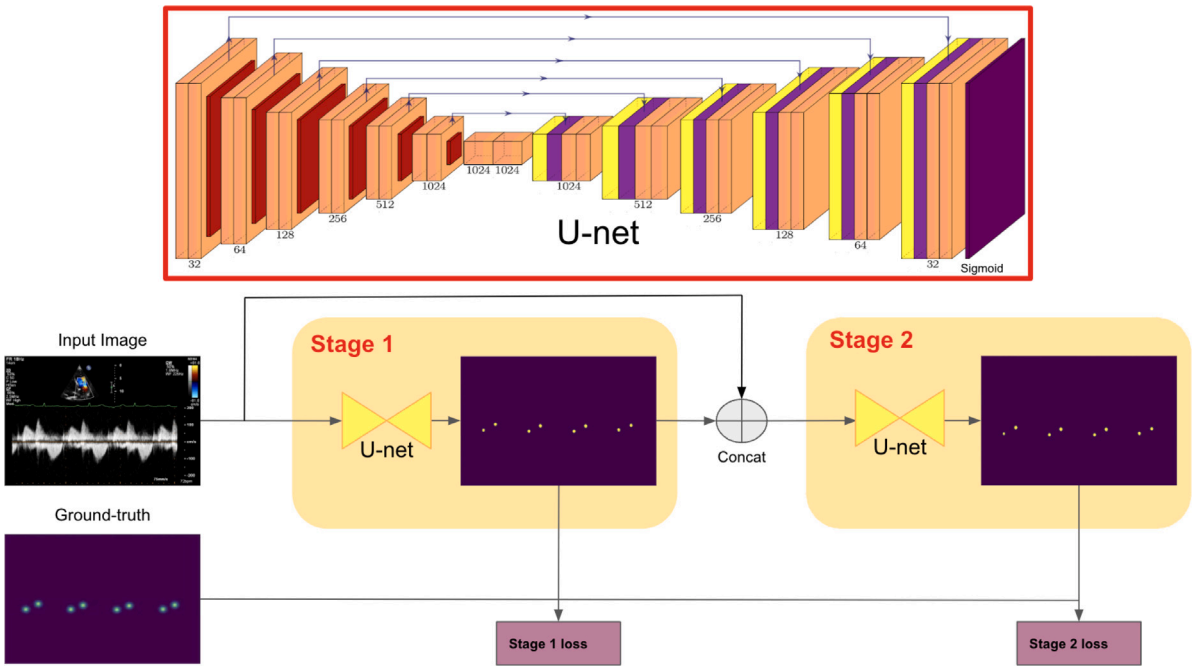

Fig. 4. Illustration of the proposed network. The backbone at each stage is U-NET [27] with the depth of 6. The U-net illustration was created using the PlotNeuralNet tool (https://github.com/HarisIqbal88/PlotNeuralNet).

2.3. Deep learning framework

In our study, we employed a heatmap regression network for keypoint detection, focusing specifically on the identification of Doppler mitral peaks. This network generates a heatmap for each image, where the intensity of each pixel represents the probability of a keypoint being present at that particular location [28,29].

Heatmap regression networks leverage the spatial distribution of features within an image to predict the likelihood of keypoints' locations. This approach is formalised by the function $H(x, y) = P(k \mid x, y)$, where H is the generated heatmap, P is the probability, and k is the presence of a keypoint at location (x, y).

The network architecture is designed to optimise this probability distribution across the image, ensuring that the peaks in the heatmap align with the actual keypoints in the image, such as the E-and A-waves in Doppler mitral inflow images. This method allows for precise localisation of keypoints, even in complex medical images where the keypoints of interest may vary in number and position due to patient-specific factors or the quality of the imaging process.

Fig. 4. provides a schematic overview of the proposed automatic measurement framework, which is comprised of a multi-stage network in which the output from each stage is concatenated with the original image and sent to the next stage [30–32].

By delivering heatmap estimations along with the original image to each successive stage, the network can reassess its initial predictions. These heatmaps function as a versatile, non-parametric representation of the spatial uncertainty associated with each keypoint location. This grants the subsequent network stage with invaluable information, empowering it to develop rich, image-dependent spatial models that capture the relationships between the keypoints. In this process, the model learns from its previous attempts, especially focusing on parts of the image where it was less certain, to improve its predictions.

Based on research done by Newell et al. [30], we applied the loss at each stage and optimised with the same Ground-truth. In our study, Binary Cross Entropy and Sørensen Dice Coefficient were minimised in unison. Kurmann et al. [33] demonstrated that Cross-Entropy loss is a feasible option for landmark detection tasks. The use of Dice loss can facilitate the detection of multiple keypoints on a single heatmap by a model.

The Dice loss formula is shown in Eq. (2).

$$1 - \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2}$$

Here, A represents the Ground-truth and B represents the prediction. The formula will compute how much the predicted keypoint heatmap overlaps with its Ground-truth. Binary Cross-Entropy can be defined in Eq. (3).

$$\frac{1}{N} \sum_{i=1}^{N} -(Y_i * \log(P_i) + (1 - Y_i) * (\log(1 - P_i))) \tag{3}$$

Here, P is a predicted heatmap, Y is a Ground-Truth representation and N is number of predicted samples.

Compared to methods, which predict a fixed number of keypoints, and each keypoint is predicted on a separate channel, our model can predict a varying number of keypoints on the same channel.

Empirically, we employed a two-stage network, and at each stage we implemented a U-Net-like [27] architecture with an encoder–decoder structure. Each stage receives an input of shape 1024×1024 with N channels (although we defined input images for the model to be of Greyscale with one channel, stages can process multi-channel matrices, which allows the concatenation of original image with previous stage output), and produces an output of the same shape, 1024×1024, but with 2 channels; each responsible for predicting peaks of one of wave type (i.e., E and A).

The input image is convolved through the first stage, which produces a set of heatmap approximations which are concatenated with the original image and employed as input for the second stage. The final set of heatmaps utilised for evaluation are generated by the second stage.

Doppler mitral inflow images can contain a varying number of heartbeats; this will depend on the patient's heart rate and the sweep speed selected by the operator during image acquisition. Therefore, our network's ability to detect an arbitrary number of peak locations in the images is crucial.

2.4. Implementation and evaluation details

The models were implemented using the TensorFlow 2.0 deep learning framework [34] and trained using an Nvidia GeForce RTX 4090.

Variable-sized images were zero-padded to a uniform size of 1024 × 1024 pixels. The model was trained using 957 images, and the parameters were fine-tuned using a validation set of 107 images, utilising a 90:10 dataset split. Training was conducted over 25 epochs with a batch size of 2. To minimise over-fitting, early stopping was employed, when training was maintained until the validation loss plateaued, restoring best weights after training stopped. The two-stage heatmap regression model was trained using the ADAM optimiser [35] with a learning rate of 0.0001, and Binary Cross Entropy + Sørensen Dice Coefficient loss. The performance of the fine-tuned model was then examined using the test dataset containing 200 images.

During post-processing, heatmap predictions, which ranged from 0 to 1, were extracted. To ensure high accuracy, we only considered those with a confidence threshold of >0.7. Although our model did not generate Gaussian-like heatmaps, values close to 1 indicated high certainty about a keypoint's location. We attribute this to our use of the Sørensen Dice Coefficient in the training loss function.

While Stern et al. [36] suggest converting the heatmap to a binary mask through thresholding and locating the mass centroid to determine the ideal keypoint position, our model tends to indicate a far higher likelihood at a point of interest. As a result, we were able to identify local maxima with a high degree of reliability without needing to locate the mass centroid.

To assess the performance of the trained deep learning model, predicted Cartesian coordinates for E- and A-wave velocities were compared to manual annotations provided by the human experts. To this end, all coordinates were converted into cm/s using OCR technique.

To obtain Cartesian coordinates from the predicted heatmaps, all local maxima were identified using a Greedy iterative process, which is described in Algorithm 1.

This strategy allows finding all significant peaks in a heatmap. However, it also has some potential issues to be aware of. Specifically, the choice of how large a region to zero out around each maximum can have a significant effect on final results. If the region is too large, there is a risk to eliminate potential peak predictions; if it is too small, the same peak can be detected multiple times.

Despite these potential issues, it still suits primary use-case, because peaks on mitral Doppler images are well separated from each other.

Subsequently, the task at hand involved locating the corresponding Ground-Truth peak point for each predicted keypoint. In order to accomplish this, a 60-pixel-wide zone centred around each Ground-Truth keypoint was defined, extending 30 pixels in both directions. This arrangement was found to be adequate in encapsulating an entire wave in virtually all instances, predicated on a commonly employed sweep speed of 100 mm/s.

A keypoint detection is assumed successful if the prediction and Ground-truth are less than 30 pixels apart on X-axis. Therefore, if a ground-truth keypoint coordinate fell within this boundary, we considered the predicted keypoint to be a successful detection (i.e., True positive). Conversely, predicted keypoints that could not be matched any Ground-truth point were assumed incorrect detections (i.e., False positive). If a Ground-truth keypoint had no corresponding prediction, it was classified as a missed keypoint (i.e., a False negative). Whole process can be summarised as follows:

1. **True Positives (TP):** These represent keypoints in the predictions set that correspond to a keypoint in the Ground-Truth set within the distance threshold.
2. **False Positives (FP):** These represent keypoints in the predictions set that lack a corresponding keypoint in the Ground-Truth set within the distance threshold.
3. **False Negatives (FN):** These represent keypoints in the Ground-Truth set that do not have a corresponding keypoint in the predictions set within the distance threshold.

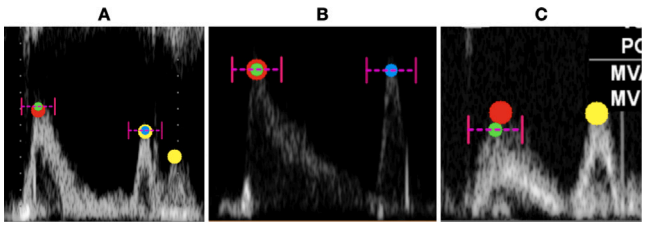

Fig. 5. Illustration of beat-matching conditions with predicted E- and A-wave coordinates (large red and yellow circles, respectively), corresponding ground-truth keypoints (small green and blue circles, respectively) and defined 30 pixels boundaries for each Ground-Truth peak (magenta coloured vertical lines). (A) Two predictions fall within Ground-Truth boundaries, resulting in True Positive predictions, and one prediction is not matched to any Ground-Truth, resulting in False Positive case. (B) E-wave prediction is considered as True Positive, and as no predicted peaks present for A-wave annotation, it is counted as False Negative case. (C) Only E-wave is counted as True Positive, whereas A-wave prediction is counted as False Positive.

Algorithm 1 Find Cartesian Coordinates from Predicted Heatmaps

Initialize: $threshold$, $distance = 30$, $max = 0$
Initialize: List to store maxima coordinates
while true **do**
  Find the global maximum: Search for the highest value in the heatmap
  for $i = 0$ to $rows - 1$ **do**
    for $j = 0$ to $columns - 1$ **do**
      if $heatmap[i][j] > max$ **then**
        $max = heatmap[i][j]$
        $x_{max} = i$
        $y_{max} = j$
      end if
    end for
  end for
  if $max < threshold$ **then**
    Terminate: Stop if the maximum is below the threshold
    break
  end if
  Store the maximum: $store(x_{max}, y_{max})$
  Zero out a region around the maximum: Clear an area around the found maximum
  for $i = x_{max} - distance$ to $x_{max} + distance$ **do**
    for $j = y_{max} - distance$ to $y_{max} + distance$ **do**
      $heatmap[i][j] = 0$
    end for
  end for
  Reset: $max = 0$ Reset the maximum for the next iteration
end while
Return: List of all stored maxima

Precision and sensitivity were computed over the pool of all heartbeats, across all patients. F1-score was calculated as the harmonic mean of precision and sensitivity.

For the 'True positive' cases, statistical analysis of the levels of agreement between the automated measurements and the human experts was performed using Bland–Altman plots; bias (mean of differences) and Standard Deviation were calculated where the confidence interval was defined as ± 1.96 SD.

An illustration of beat-matching conditions, providing examples of three possible scenarios (i.e., True positive, False positive, and False negative) are provided in Fig. 5.

3. Results and discussion

Examples of successful automated predictions (true positives) are shown in Fig. 6. The results show that the model is capable of recognising various potential appearances of mitral inflow Doppler waveforms, such as double peaks (clearly separated E and A waves), singular

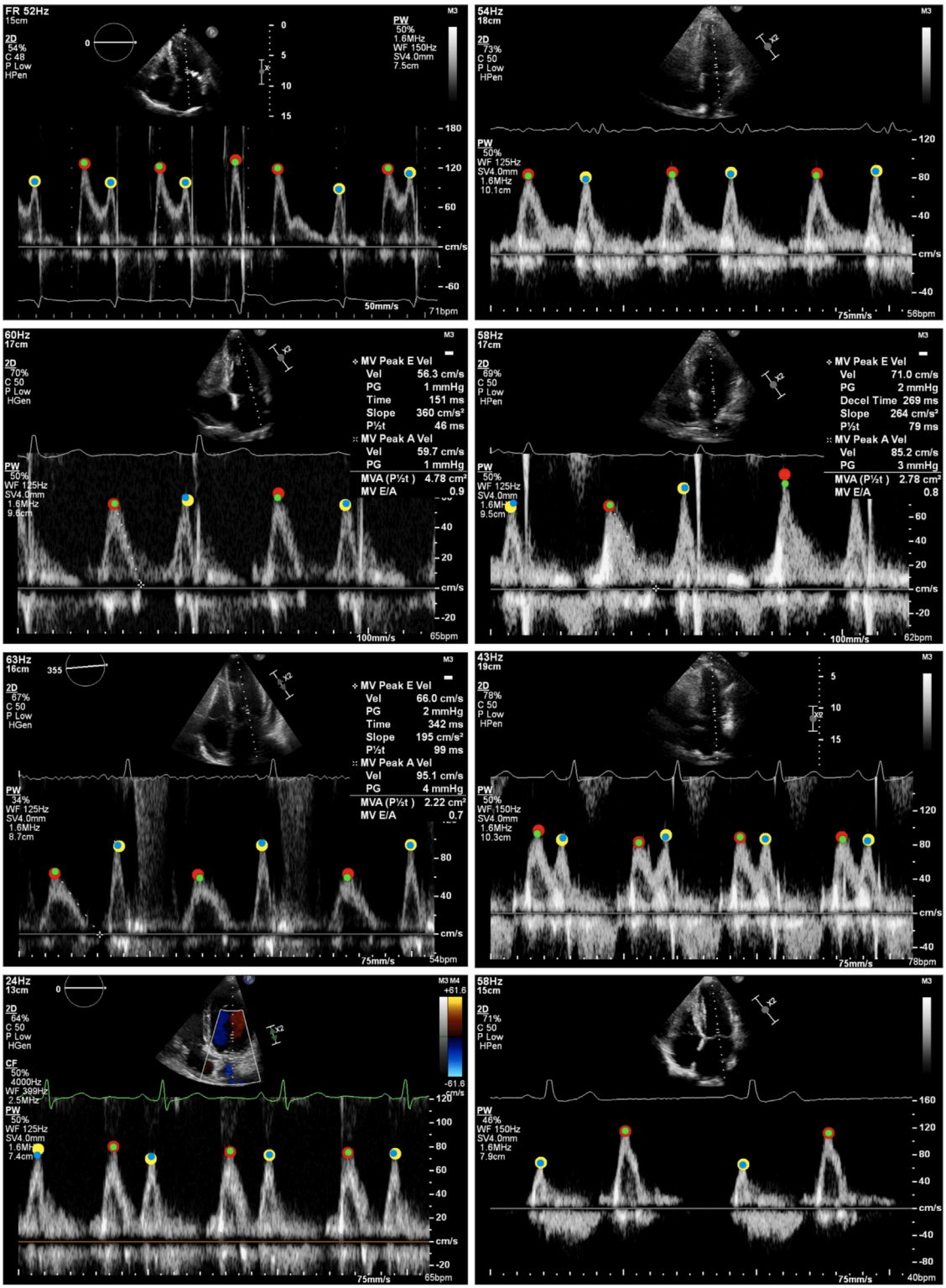

Fig. 6. Examples of successful automated predictions (true positives), showing predicted E- and A-wave coordinates (large red and yellow circles, respectively) and corresponding ground-truth keypoints (small green and blue circles, respectively).

peaks (only one wave present), overlapping waves, and peaks in noisy heartbeats.

The agreement between the model's predictions and the consensus measurements for all E- and A-waves was examined via the Bland–Altman analysis, shown in Fig. 7.

The analysis consisted of two parts. First, we undertook a beat-by-beat comparison, where the individual cardiac cycles, originating from all Doppler strips (i.e., patients) in the testing dataset, were placed in a pool of heartbeats and used for analysis.

Our Bland–Altman analysis revealed a bias of 0.29 cm/s and a standard deviation of 3.61 cm/s for the E-wave, and a bias of 0.13 cm/s with a standard deviation of 2.12 cm/s for the A-wave. When compared with the findings of Elwazir et al. [17], who reported a higher bias of 6 cm/s for the E-wave and 5 cm/s for the A-wave, along with a standard

Table 1
Performance comparison of the standard U-Net model, different U-Net backbones (ResNet-50, MobileNet v2, EfficientNet-B0), and LinkNet, in the context of detecting E-wave and A-wave keypoints. It includes an evaluation of bias and standard deviation (Std) for each model. Best value in each column is highlighted. All values are converted to cm/s.

Model	E-wave				A-wave			
	Each beat		Patient average		Each beat		Patient average	
	Bias	Std	Bias	Std	Bias	Std	Bias	Std
ResNet-50 U-Net	1.62	2.78	1.55	**1.86**	0.55	2.44	0.5	1.86
MobileNet v2 U-Net	0.45	3.65	0.39	2.44	−1.24	3.52	−1.55	4.24
EfficientNet-B0 U-Net	1.36	3.02	1.54	2.29	**−0.09**	5.24	0.23	3.89
LinkNet	1.18	**2.48**	1.18	1.9	0.31	4.5	0.18	2.13
U-Net (Main model)	**0.29**	3.61	**0.31**	2.	0.13	**2.12**	**0.14**	**1.54**

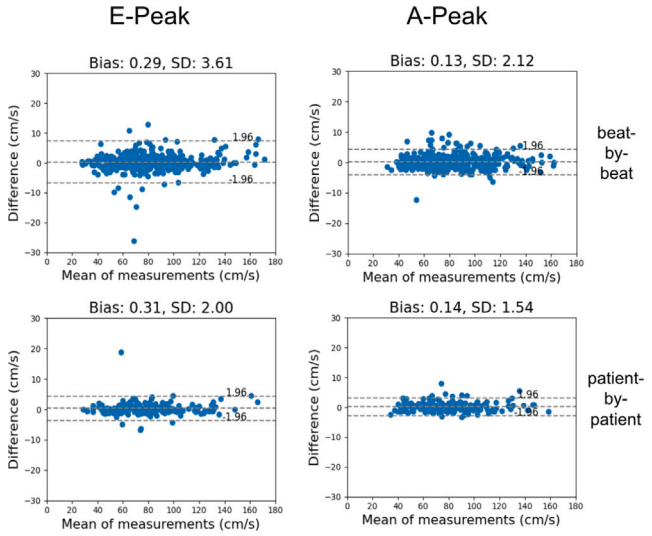

Fig. 7. Bland–Altman plots for beat-by-beat (upper row) and patient-by-patient (lower row) analysis for E (left column) and A (right column) Doppler peak velocities, where the agreement between the expert consensus and the model is shown. Beat-by-beat is the pool of all detected heartbeats present in Doppler images across all patients in the testing dataset, and patient-by-patient is when a representative measurement is obtained for each patient by taking the average of all automatically detected heartbeats (true positives) in each image.

deviation of 3 cm/s for both waves, our results indicate lower bias but higher Standard deviation. In contrast, the study by Zamzmi et al. [16] demonstrated even smaller biases of −0.6 cm/s for the E-wave and −0.7 cm/s for the A-wave. The estimated standard deviations, 1.5 cm/s for the E-wave and 2.6 cm/s for the A-wave, are somewhat comparable to ours but reflect a different pattern in measurement precision. It should be noted that their approach can be considered as semi-automated, as it required manual segmentation of heartbeats in instances where the automated process was unsuccessful.

Additionally, we extended our analysis to a patient-by-patient basis, where a representative peak velocity measurement was obtained for each Doppler image (i.e., patient) in the testing dataset by taking average of measurements across all heartbeats present in that image. The patient-by-patient analysis yielded a bias of 0.31 cm/s and a standard deviation of 2.00 cm/s for the E-wave. For the A-wave, the bias was at 0.14 cm/s, with a lower standard deviation of 1.54 cm/s.

A slightly better patient-by-patient agreement between the two methods may be due to averaging multiple cardiac cycles in a Doppler strip that can potentially reduce the effect of potential outliers.

No systematic bias was observed in either analysis. The results underscore the model's ability to capture and predict the expert consensus, thus reinforcing the potential of our model in the realm of automated peak velocity measurements in mitral inflow Doppler images.

While the U-Net architecture served as our primary model, we also examined the performance of alternative network architectures as our backbone models for keypoint detection, specifically ResNet-50 [37], MobileNet v2 [38], EfficientNet-B0 [39], and LinkNet [40], all implemented using the Segmentation Models library [41]. For each model, we measured bias and standard deviation for both E-wave and A-wave keypoints, considering the metrics for each beat (beat-by-bet) and patient average (patient-by-patient).

While this comparative analysis yielded valuable insights into the different architectural approaches, the results, presented in Table 1, were clear: U-Net consistently maintained strong performance across the majority of measurements. No single alternative architecture demonstrated a definitive advantage over U-Net. Considering its strong performance and established position as a well-suited model for medical image processing, we opted for U-Net as the core model for our subsequent investigations.

3.1. Inter- and intra-observer variability

Whereas previous evaluations have shed light on our model's alignment with the consensus of expert annotations, they do not fully capture the extent of variability among the experts themselves.

Therefore, the measurements by individual experts were compared with the consensus (mean) measurement. The standard deviation of the pooled data from all five experts for all heartbeats, indicating the inter-observer variability, was 3.25 cm/s and 2.79 cm/s for E- and A-waves, respectively. This measure demonstrate the inherent subjectivity in the task of annotating peak velocities in mitral inflow Doppler images, thereby providing a context within which the performance of our automated system can be assessed.

Using a common reference (average of all experts) unfairly favour the human experts because they are part of the reference; the common reference could not be considered independent from the expert under study. To ensure fair comparison, Fig. 8 plots agreements between two sets of measurements; each human expert is compared with 4 other experts, when their consensus (mean) is considered as the reference annotation (blue boxplots). The model is also compared with the consensus of the same 4 human annotations (orange boxplots).

All 10 panels suggest performance of the model is similar, if not better, to that of an individual expert when using the other experts as a reference standard.

Since different experts make different judgments, it is not possible for any automated model to agree with all experts. However, it is desirable for the model to not be an outlier when compared with the distribution of human judgments; that is, to behave approximately as well as a human expert.

To test this, a simple visual summary of the automated measurements in the context of expert performance is given in Fig. 9 for the E-wave velocity, where the model's predictions for each heartbeat are represented by green dots, and individual human measurements are denoted by grey dots.

A close observation of the figure reveals an interesting pattern: the model's measurements consistently fall within the spread and range of the experts' annotations. This result implies that our model's predictions are not outliers, but rather align with the variability that

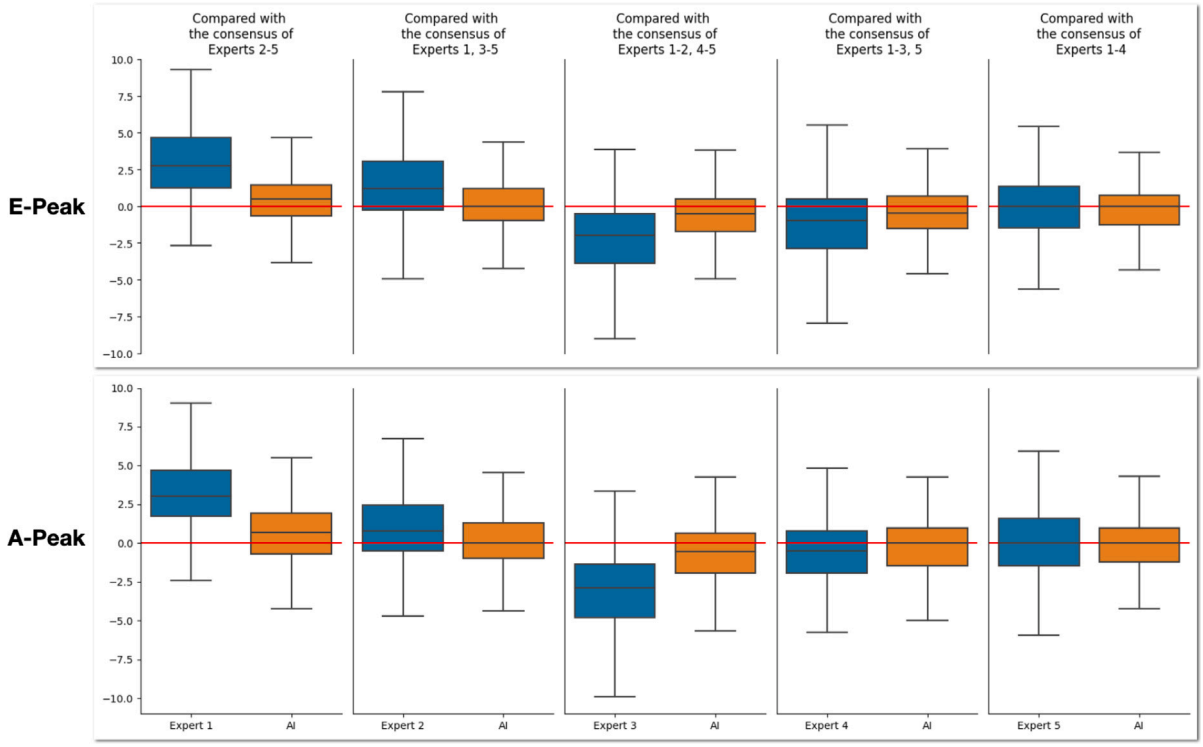

Fig. 8. Comparative Analysis of human expert and AI evaluations for E-wave (upper panel) and A-wave (lower panel) velocity values. Each human expert is compared to the consensus (mean) of all other 4 experts (blue boxplots). In each case, alongside these comparisons, are those obtained from the AI model relative to the consensus of the same 4 annotations (orange boxplots). In the box-and-whisker plots, the thick line represents the median, the box represents the quartiles, and the whiskers represent the 2.5% and 97.5% percentiles.

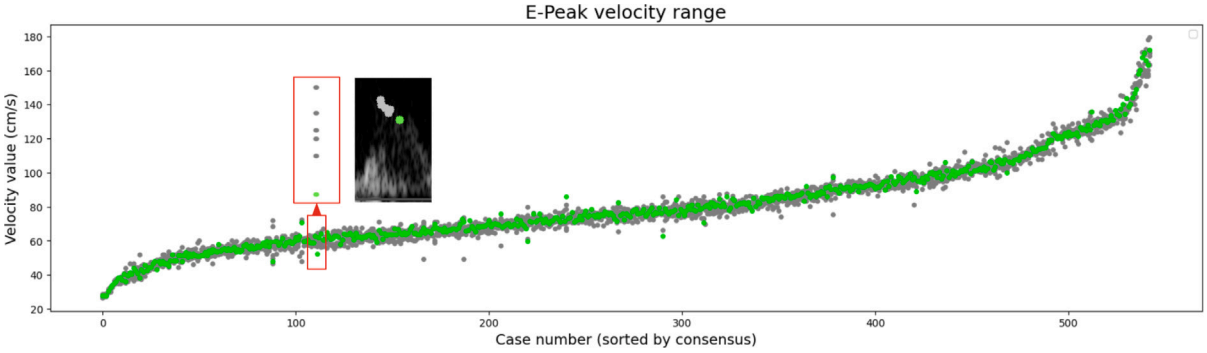

Fig. 9. Performance of the automated model in quantifying E-wave Doppler velocities (green dots) in the context of range of manual measurements by individual human experts (grey dots) across all heartbeats in the testing dataset. The measurements are presented in ascending order of peak velocity, as defined by expert consensus. Also shown, is the magnified version of an outlier, when the automated prediction fell outside the manual range, together with the corresponding Doppler wave.

naturally arises among the expert annotators. In other words, our model's performance mirrors the human experts' range of agreement, effectively positioning its predictions within the same range (or close to it) of variability, that characterises expert human annotations. The A-wave velocity measurements had a similar pattern.

The range of human expert judgments for each heartbeat may be assumed as the uncertainty of the reference method and, therefore, the highest accuracy obtainable. The mean velocity range was 6.58 ± 3.86 cm/s and 6.87 ± 3.49 cm/s for E- and A-wave, respectively.

For each heartbeat, there were six measurements for each velocity peak (five human and one automated). By chance alone, in one-third (33.3%) of the cases, the measurement of an individual "operator" (human or automated) would be the smallest or the largest among the six measurements (one-sixth chance of being smallest + one-sixth chance of being largest).

As shown in Fig. 10, the model performs similarly to human operators: it is an outlier sometimes, but so is each of the humans. For E-wave velocities, expert 1 had the highest percentage of 48.3% for being at an extreme. Expert 3 had the second highest percentage (43.0%) of heartbeats for being the outlier. For A-wave velocities, expert 3 had the highest percentage of 56.6% for being at an extreme.

The model was the outlier in only 9.6% and 9.0% of the heartbeats for E- and A-waves, respectively. This suggests that the model performed no worse than human experts in measuring peak velocities.

The intra-observer variability, representing the mean difference between two measurements by the same expert on separate occasions, was 1.71 ± 4.32 cm/s and 1.55 ± 2.45 cm/s for E-wave and A-wave velocities, respectively. A second expert, repeating measurements twice, showed similar variability, with values of 0.09 ± 2.81 cm/s for E-wave velocities and 0.51 ± 2.8 cm/s for A-wave velocities.

Table 2
Precision, Recall and F1 Score for detection of peak velocity waves by each human expert and the automated model, when compared to the corresponding independent consensus of human experts. The performance of the model has also been provided compared to the consensus of all human experts.

	Reference measurement	E-wave			A-wave		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Expert 1	Consensus of experts 2–5	0.98	0.97	0.98	0.97	0.98	0.97
Model		0.96	0.91	0.93	0.96	0.88	0.91
Expert 2	Consensus of experts 1, 3–5	0.93	1.00	0.96	0.93	0.99	0.96
Model		0.95	0.91	0.93	0.95	0.88	0.91
Expert 3	Consensus of experts 1–2, 4–5	0.95	0.98	0.96	0.94	0.98	0.96
Model		0.95	0.91	0.93	0.96	0.87	0.91
Expert 4	Consensus of experts 1–3, 5	0.99	0.98	0.98	0.93	0.99	0.96
Model		0.96	0.91	0.93	0.95	0.88	0.91
Expert 5	Consensus of experts 1–4	0.98	0.95	0.96	0.99	0.93	0.96
Model		0.96	0.91	0.93	0.96	0.87	0.91
Model	**Consensus of experts 1–5**	**0.96**	**0.91**	**0.93**	**0.96**	**0.87**	**0.91**

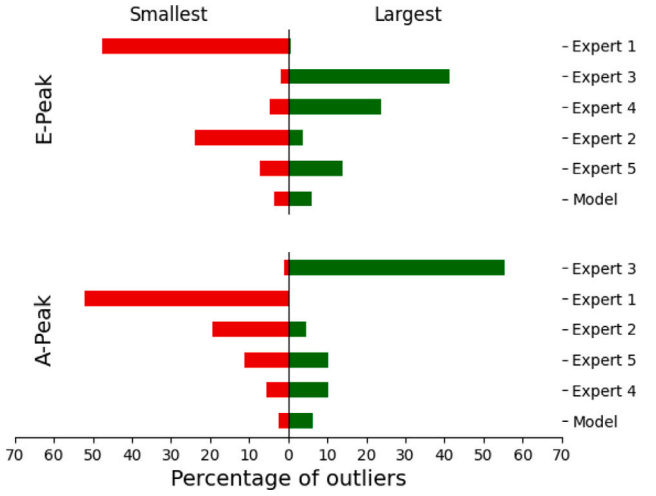

Fig. 10. Relative frequency of being the smallest (left side) or largest (right side) peak Doppler velocity measurement for the E-wave (upper panel) and A-wave (lower panel). All 6 operators (5 human experts and one automated model) are sorted based on the frequency of being the extreme (largest or smallest).

3.2. Detection capacity

Table 2 presents the Precision, Recall, and F1-score metrics for all comparisons. It is evident that the model demonstrates a similar precision, but slightly lower recall compared to individual experts when it comes to detecting and classifying Doppler peak velocities. Human experts had precision and recall of ≥93% in detecting both E- and A-wave Doppler peaks. The model had precision and recall of ≥95% and ≥87%, respectively.

When compared to consensus of all human experts, the automated model had precision 96% in detecting both types Doppler peaks, implying instances of undetected/missed or misclassified peak points.

The model showed a lower recall of 87% in detecting A-waves, compared to 91% for E-waves. Lower recall suggests the model may misclassify peaks or incorrectly identify artifacts as peak points. These findings highlight potential areas for further improvement in our model. By focusing on strategies to enhance recall without significantly impacting precision, the model's overall performance could be further optimised.

Fig. 11. depicts examples of failed predictions. Our observations underscore that a primary obstacle for our neural network model involves addressing peaks surrounded by, or located adjacent to, artifacts. This is manifested by several missed A-wave points it the figure. Such circumstances significantly complicate the model's capacity to discriminate between an actual artifact and a relevant peak. In such instances, it falls upon the human experts to leverage their expertise and experience to discern and eliminate the artifact signal during the processing stage. Additionally, our model encounters difficulties when faced with peaks that extensively overlap, thereby intensifying the task of differentiating between them.

Another complication arises when dealing with peaks exhibiting low contrast which, due to their subdued visibility, present a challenge for the model to detect. The model also grapples with occasional anomalies where it incorrectly identifies an irrelevant location for a peak or misclassifies it.

Despite the apparent clarity of these situations to the human eye, they continue to pose significant challenges for the models. This highlights the complex nuances of image interpretation that our model must learn to master.

In order to examine the impact of these occurrences on the final reported measurements for each patient, we conducted another patient-by-patient Bland–Altman analysis, focusing on representative values obtained from the automated model. Rather than solely comparing true positive cases in a pairwise manner, we independently calculated the average velocity for the manual and automated measurements for each individual patient, and then compared the average velocities.

For instance, consider a scenario for a single patient, where the human experts identified 6 E-waves (i.e., heartbeats), and we computed their average velocity. Meanwhile, the model correctly detected 4 E-waves (true positives), missed two E-waves (false negatives), and incorrectly identified an artifact as an E-wave point (false positive). As a result, the average automated E-wave measurement for this patient was determined from 5 predictions, representing what the model would report.

Subsequently, we compared these representative automated and manual measurements, as depicted in Fig. 12. There is a slight increase in the discordance between the model and consensus measurements compared to what was presented in Fig. 7, which only encompassed true positive cases. Nevertheless, this difference still falls within the range of inter- and intra-observer variability. This clearly demonstrates the feasibility of using our proposed model to reliably detect and measure peak Doppler velocities, independent of ECG information.

From the complete test dataset, only one image (0.5%) was found for which the model failed to detect any heartbeat. This image is showcased in the bottom-right corner of Fig. 11. The likely cause is the presence of numerous heartbeats in the Doppler strip, a result of the sweep speed chosen by the operator during image acquisition. The model encountered a singular example in the testing dataset, and there were no analogous instances in the development dataset from which it could learn.

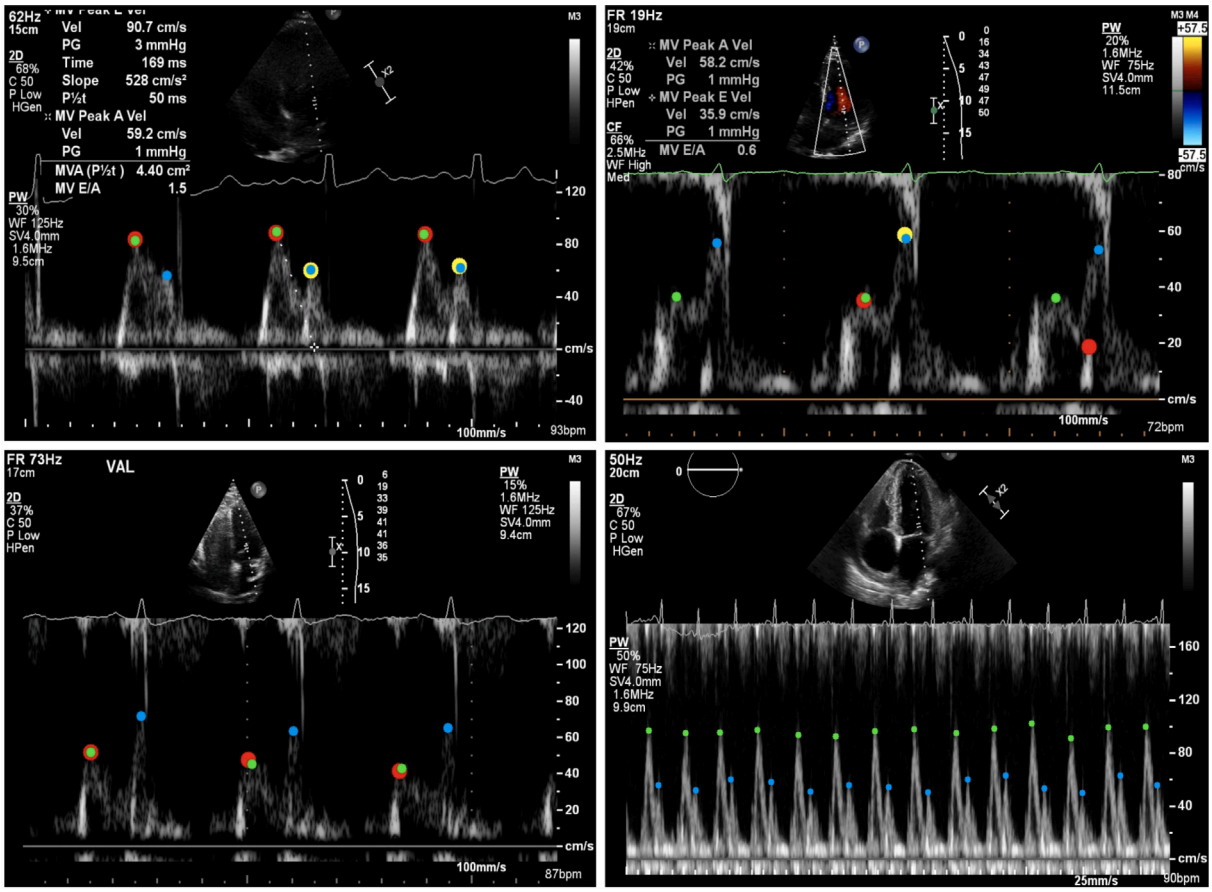

Fig. 11. Examples of failed predictions, with automated E- and A-wave predictions (large red and yellow circles, respectively) and ground-truth E- and A-wave annotations (small green and light blue circles, respectively). Top left: one missed A-wave (false negative). Top right: 4 missed waves (false negative) and one artifact peak (false positive). Bottom left: 3 missed A-waves (false negatives). Bottom right: missed all measurements for this patient.

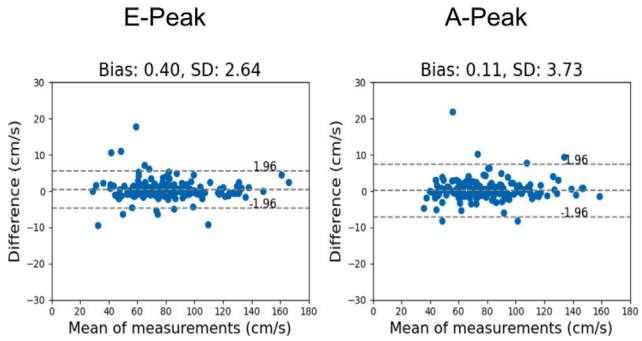

Fig. 12. As in Fig. 7 (lower row), but automated values include both true and false positives to represent what the model would report for a patient, if deployed in the clinics.

4. Conclusions

This study investigates the feasibility of using deep neural networks for fully automated detection and measurement of peak velocities in mitral Doppler inflow in 2D echocardiography, and independent from the ECG signal.

The automated model were successful in detecting E- and A-wave velocities in 96% of the heartbeats. The performance of the model is similar to that of human experts. Experts do not completely agree on where to measure the peak Doppler velocity because the judgment is complex, but the automated model behaves similarly to human experts, being no more likely to be an outlier than the experts. We believe that performing as well as a human operator indicates reasonable performance of an automated algorithm.

The model discrepancy with the expert consensus falls comfortably within the range of inter- and intra-observer variability, which underscore the inherent variability between and even within individual expert annotations, reflecting the complex nature of this task. This demonstrates the reliability of the model in measuring the peak Doppler velocities, compared with the experienced human experts.

The assessment of cardiac timing and heartbeat detection in an echocardiogram examination relies on analysing an accompanying ECG signal. However, the ECG recordings often involve the cumbersome and occasionally inconvenient setup of multiple cables. In an age where highly portable scanners can conduct targeted studies lasting just a few minutes [42], the ability to detect cardiac timing events autonomously, without relying on the ECG signal, holds significant potential for integrating automated technology into handheld devices.

In the absence of publicly available echocardiography datasets specifically tailored for mitral Doppler inflow measurements and corresponding gold-standard expert annotations, we have made our patient dataset accessible for download at intsav.github.io/doppler.html [23]. By doing so, we aim to establish a benchmark for future studies and foster advancements in this field. Additionally, we have made all developed models from this study available under open-source agreements, inviting others to scrutinise, adapt, and enhance them. This transparency in sharing both datasets and models encourages external validation of our findings, ensuring the robustness and applicability of our automated approach in clinical practice and research.

4.1. Future work

Building upon the current success of the deep learning model for automated peak velocity detection, several exciting avenues exist for future research and development. These advancements hold the potential to further refine the model's performance, broaden its clinical applications, and ultimately improve patient care.

One promising direction involves augmenting the training dataset with synthetic cases generated by Generative Adversarial Networks (GANs) [43]. This would allow the model to encounter specific scenarios that are challenging or underrepresented in real-world data, such as complex valvular abnormalities or irregular rhythms. By strengthening its ability to handle diverse situations, the model's generalisability and robustness would be significantly enhanced.

Another intriguing path lies in utilising self-supervised learning techniques to pre-train the model on unlabelled data [44]. This approach would leverage the vast wealth of readily available echocardiograms, even those without manually labelled measurements, to improve the model's initial performance and learning efficiency. By extracting valuable features and representations from unlabelled data, the model would be significantly better prepared for fine-tuning with labelled data, ultimately leading to a more robust and generalisable solution.

Furthermore, expanding the dataset to include a more diverse range of cases could further enhance the model's ability to handle variations in patient demographics, pathologies, and imaging conditions. This expansion would contribute to a more comprehensive and representative training dataset, improving the model's real-world applicability.

Exploring the feasibility of real-time application is crucial for integrating the automated model into clinical workflows. Optimising the model's efficiency and speed could enable its use in real-time echocardiography examinations, providing timely and accurate assessments during patient evaluations.

Finally, considering the addition of other Doppler measurements, such as the left ventricular outflow tract, could broaden the scope of the model's applications. Incorporating multiple Doppler measurements may enable a more comprehensive assessment of cardiac function, offering valuable insights into different aspects of cardiovascular health.

CRediT authorship contribution statement

Jevgeni Jevsikov: Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Conceptualization. **Tiffany Ng:** Data curation. **Elisabeth S. Lane:** Writing – review & editing. **Eman Alajrami:** Writing – review & editing. **Preshen Naidoo:** Writing – review & editing. **Patricia Fernandes:** Writing – review & editing. **Joban S. Sehmi:** Data curation. **Maysaa Alzetani:** Data curation. **Camelia D. Demetrescu:** Data curation. **Neda Azarmehr:** Writing – review & editing. **Nasim Dadashi Serej:** Supervision. **Catherine C. Stowell:** Data curation. **Matthew J. Shun-Shin:** Data curation. **Darrel P. Francis:** Data curation. **Massoud Zolgharni:** Conceptualization, Data curation, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research and open-access release of the has been conducted under: The Imperial College London and University of West London, United Kingdom [IRAS: 279328, REC:20/SC/0386]. This work was supported in part by the British Heart Foundation, UK (Grant no. RG/F/22/110059). J. Jevsikov is supported by the Vice Chancellor's Scholarship at the University of West London.

References

[1] Marc M. Corriveau, K. Wayne Johnston, Interobserver variability of carotid Doppler peak velocity measurements among technologists in an ICAVL-accredited vascular laboratory, J. Vasc. Surg. 39 (4) (2004) 735–741.
[2] Niti M Dhutia, Massoud Zolgharni, Michael Mielewczik, Madalina Negoita, Stefania Sacchi, Karikaran Manoharan, Darrel P Francis, Graham D Cole, Open-source, vendor-independent, automated multi-beat tissue Doppler echocardiography analysis, Int. J. Cardiovasc. Imaging 33 (2017) 1135–1148.
[3] Massoud Zolgharni, Niti M Dhutia, Graham D Cole, M Reza Bahmanyar, Siana Jones, SM Afzal Sohaib, Sarah B Tai, Keith Willson, Judith A Finegold, Darrel P Francis, Automated aortic Doppler flow tracing for reproducible research and clinical measurements, IEEE Trans. Med. Imaging 33 (5) (2014) 1071–1082.
[4] Elaine YL Lui, Aaron H Steinman, Richard SC Cobbold, K Wayne Johnston, Human factors as a source of error in peak Doppler velocity measurement, J. Vasc. Surg. 42 (5) (2005) 972–e1.
[5] Sherif F Nagueh, Otto A Smiseth, Christopher P Appleton, Benjamin F Byrd, Hisham Dokainish, Thor Edvardsen, Frank A Flachskampf, Thierry C Gillebert, Allan L Klein, Patrizio Lancellotti, et al., Recommendations for the evaluation of left ventricular diastolic function by echocardiography: an update from the American society of echocardiography and the European association of cardiovascular imaging, Eur. J. Echocardiogr. 17 (12) (2016) 1321–1360.
[6] Massoud Zolgharni, Niti M Dhutia, Graham D Cole, Keith Willson, Darrel P Francis, Feasibility of using a reliable automated Doppler flow velocity measurements for research and clinical practices, in: Medical Imaging 2014: Ultrasonic Imaging and Tomography, Vol. 9040, SPIE, 2014, pp. 360–368.
[7] Andrew F. Hall, Scott P. Nudelman, Sándor J. Kovács, Beat averaging alternatives for transmitral Doppler flow velocity images, Ultras. Med. Biol. 24 (7) (1998) 971–979.
[8] JinHyeong Park, S Kevin Zhou, John Jackson, Dorin Comaniciu, Automatic mitral valve inflow measurements from Doppler echocardiography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2008, pp. 983–990.
[9] Emmanuel Gaillard, Lyes Kadem, Marie-Annick Clavel, Philippe Pibarot, Louis-Gilles Durand, Optimization of Doppler echocardiographic velocity measurements using an automatic contour detection method, Ultras. Med. Biol. 36 (9) (2010) 1513–1524.
[10] Amirtahà Taebi, Richard H Sandler, Bahram Kakavand, Hansen A Mansy, Estimating peak velocity profiles from doppler echocardiography using digital image processing, in: 2018 IEEE Signal Processing in Medicine and Biology Symposium, SPMB, IEEE, 2018, pp. 1–4.
[11] N.V. Kiruthika, B. Prabhakar, M. Ramasubba Reddy, Automated assessment of aortic regurgitation using 2D Doppler echocardiogram, in: Proceedings of the 2006 IEEE International Workshop on Imagining Systems and Techniques (IST 2006), IEEE, 2006, pp. 95–99.
[12] Nagashettappa Biradar, M.L. Dewal, Manoj Kumar Rohit, Automated delineation of Doppler echocardiographic images using texture filters, in: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2015, pp. 1903–1907.
[13] T Syeda-Mahmood, P Turaga, David Beymer, Fei Wang, Arnon Amir, Hayit Greenspan, K Pohl, Shape-based similarity retrieval of Doppler images for clinical decision support, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 855–862.
[14] Hayit Greenspan, Oron Shechner, Mickey Scheinowitz, Micha S Feinberg, Doppler echocardiography flow-velocity image analysis for patients with atrial fibrillation, Ultras. Med. Biol. 31 (8) (2005) 1031–1040.
[15] O. Shechner, M. Scheinowitz, M.S. Feinberg, H. Greenspan, Automated method for doppler echocardiography image analysis, in: 2004 23rd IEEE Convention of Electrical and Electronics Engineers in Israel, IEEE, 2004, pp. 177–180.
[16] Ghada Zamzmi, Li-Yueh Hsu, Wen Li, Vandana Sachdev, Sameer Antani, Fully automated spectral envelope and peak velocity detection from Doppler echocardiography images, in: Medical Imaging 2020: Computer-Aided Diagnosis, Vol. 11314, SPIE, 2020, pp. 1053–1064.
[17] Mohamed Y Elwazir, Zeynettin Akkus, Didem Oguz, Zi Ye, Jae K Oh, Fully automated mitral inflow doppler analysis using deep learning, in: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2020, pp. 691–696.
[18] Tollef Struksnes Jahren, Erik N Steen, Svein Arne Aase, Anne H Schistad Solberg, Estimation of end-diastole in cardiac spectral doppler using deep learning, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (12) (2020) 2605–2614.
[19] Feifei Yang, Xiaotian Chen, Xixiang Lin, Xu Chen, Wenjun Wang, Bohan Liu, Yao Li, Haitao Pu, Liwei Zhang, Dangsheng Huang, et al., Automated analysis of doppler echocardiographic videos as a screening tool for valvular heart diseases, Cardiovasc. Imaging 15 (4) (2022) 551–563.
[20] Ke Sun, Bin Xiao, Dong Liu, Jingdong Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.
[21] Elisabeth S Lane, Jevgeni Jevsikov, Matthew J Shun-Shin, Niti Dhutia, Nasser Matoorian, Graham D Cole, Darrel P Francis, Massoud Zolgharni, Automated multi-beat tissue Doppler echocardiography analysis using deep neural networks, Med. Biol. Eng. Comput. (2023) 1–16.

[22] Neda Azarmehr, Xujiong Ye, James P Howard, Elisabeth S Lane, Robert Labs, Matthew J Shun-Shin, Graham D Cole, Luc Bidaut, Darrel P Francis, Massoud Zolgharni, Neural architecture search of echocardiography view classifiers, J. Med. Imaging 8 (3) (2021) 034002.

[23] Intelligent sensing and vision research group (IntSaV), 2024, URL https://intsav.github.io/doppler.html. Accessed: 2024-01-23.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226–231, (34).

[25] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, Xiaowei Xu, DBSCAN revisited, revisited: why and how you should (still) use DBSCAN, ACM Trans. Database Syst. 42 (3) (2017) 1–21.

[26] Vasileios Belagiannis, Andrew Zisserman, Recurrent human pose estimation, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 468–475.

[27] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[28] Jonathan J Tompson, Arjun Jain, Yann LeCun, Christoph Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, Adv. Neural Inf. Process. Syst. 27 (2014).

[29] Tomas Pfister, James Charles, Andrew Zisserman, Flowing convnets for human pose estimation in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1913–1921.

[30] Alejandro Newell, Kaiyu Yang, Jia Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, Springer, 2016, pp. 483–499.

[31] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, Convolutional pose machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[32] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, Dimitris Metaxas, Quantized densely connected u-nets for efficient landmark localization, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 339–354.

[33] Thomas Kurmann, Pablo Marquez Neila, Xiaofei Du, Pascal Fua, Danail Stoyanov, Sebastian Wolf, Raphael Sznitman, Simultaneous recognition and pose estimation of instruments in minimally invasive surgery, in: Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20, Springer, 2017, pp. 505–513.

[34] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016, arXiv preprint arXiv:1603.04467.

[35] Diederik P. Kingma, Jimmy Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[36] Antonia Stern, Lalith Sharan, Gabriele Romano, Sven Koehler, Matthias Karck, Raffaele De Simone, Ivo Wolf, Sandy Engelhardt, Heatmap-based 2d landmark detection with a varying number of landmarks, in: Bildverarbeitung FÜR Die Medizin 2021, Springer, 2021, pp. 22–27.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[39] Mingxing Tan, Quoc Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[40] Abhishek Chaurasia, Eugenio Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in: 2017 IEEE Visual Communications and Image Processing, VCIP, IEEE, 2017, pp. 1–4.

[41] Pavel Iakubovskii, Segmentation models, 2019, https://github.com/qubvel/segmentation_models.

[42] Ariane Testuz, Hajo Müller, Pierre-Frederic Keller, Philippe Meyer, Tomoe Stampfli, Lucka Sekoranja, Cedric Vuille, Haran Burri, Diagnostic accuracy of pocket-size handheld echocardiographs used by cardiologists in the acute care setting, Eur. Heart J.–Cardiovasc. Imaging 14 (1) (2013) 38–42.

[43] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[44] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, Fillia Makedon, A survey on contrastive self-supervised learning, Technologies 9 (1) (2020) 2.