User-Intended Doppler Measurement Type Prediction Combining CNNs With Smart Post-Processing

Andrew Gilbert, Marit Holden, Line Eikvil, Mariia Rakhmail, Aleksandar Babić, Svein Arne Aase, Eigil Samset, Kristin McLeod

***Abstract*— Spectral Doppler measurements are an important part of the standard echocardiographic examination. These measurements give insight into myocardial motion and blood flow, providing clinicians with parameters for diagnostic decision making. Many of these measurements are performed automatically with high accuracy, increasing the efficiency of the diagnostic pipeline. However, full automation is not yet available because the user must manually select which measurement should be performed on each image. In this work, we develop a pipeline based on convolutional neural networks (CNNs) to automatically classify the measurement type from cardiac Doppler scans. We show how the multi-modal information in each spectral Doppler recording can be combined using a meta parameter post-processing mapping scheme and heatmaps to encode coordinate locations. Additionally, we experiment with several architectures to examine the tradeoff between accuracy, speed, and memory usage for resource-constrained environments. Finally, we propose a confidence metric using the values in the last fully connected layer of the network and show that our confidence metric can prevent many misclassifications. Our algorithm enables a fully automatic pipeline from acquisition to Doppler spectrum measurements. We achieve 96% accuracy on a test set drawn from separate clinical sites, indicating that the proposed method is suitable for clinical adoption.**

***Index Terms*— Convolutional neural network (CNN), deep learning, classification, ultrasound (US), Doppler**

I. Introduction

ECHOCARDIOGRAPHY is the primary method used to image the heart due to its portability, affordability, and absence of ionizing radiation. The diagnostic power of echocardiography is reflected in clinical guidelines. Echocardiography indices are included as both minor and major clinical diagnostic criteria in many protocols [1]. As computational power increases image quality improves. Consequently, the theoretical accuracy of clinical measurements also increases.

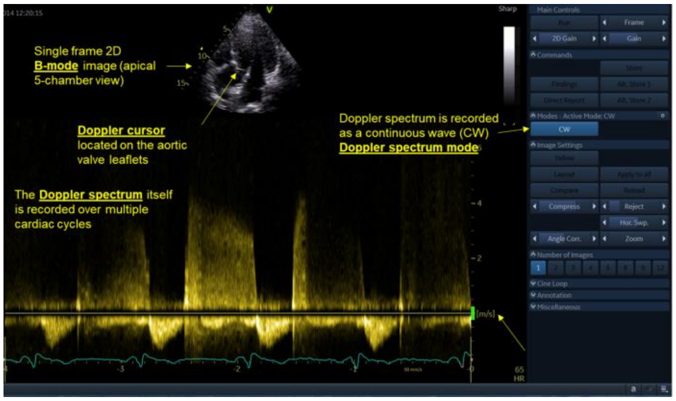

Fig. 1. Example of a Doppler acquisition shown in EchoPAC (GE Healthcare, Horten, NO) depicting the relevant information to a spectrum classification problem as a clinician would see it.

In addition to the diagnostic power, there is a growing trend to use echocardiography as a therapy guidance tool to support interventions and complement other imaging modalities. Minimally invasive valve interventions are much less risky than full surgery and are becoming the therapy of choice as techniques and prosthetics advance. Spectral Doppler imaging is the primary method to assess blood flow across valves, a crucial step for intervention planning and follow-up [2]. Therefore, spectral Doppler imaging has become an integral component of the echocardiography exam to provide a means to assess hemodynamic function in all four valves of the heart.

A. Spectral Doppler Measurements

Fig. 1 shows an example of a spectral Doppler acquisition as seen in EchoPAC (GE Healthcare, Horten, NO). There are many important features of the acquisition that are available within the raw data of each recording:

- The **Doppler spectrum** is displayed over multiple cardiac cycles for analysis and measurement.
- The **relative baseline** of the Doppler spectrum can be

This project has received funding from the European Union's Horizon 2020 research and Innovation program under the Marie Sklodowska-Curie grant agreement No 764738.

A. Gilbert and E. Samset are with GE Healthcare and also with the Department of Informatics at the University of Oslo, both Oslo, NO (email: andrew.gilbert@ge.com; eigil.samset@ge.com).

M. Holden and L. Eikvil are with the Norwegian Computing Center, Oslo, NO (email: marit.holden@nr.no; line.eikvil@nr.no).

A. Babić and M. Rakhmaiil were with GE Healthcare in Oslo, NO for their contribution to this work (email: mariya.rakhmayil@ge.com and aleksandar.babic@gmail.com).

S. A. Aase, and K. McLeod are with GE Healthcare in Oslo, NO (email: sveinarne.aase@ge.com; and kristin.mcleod@ge.com).

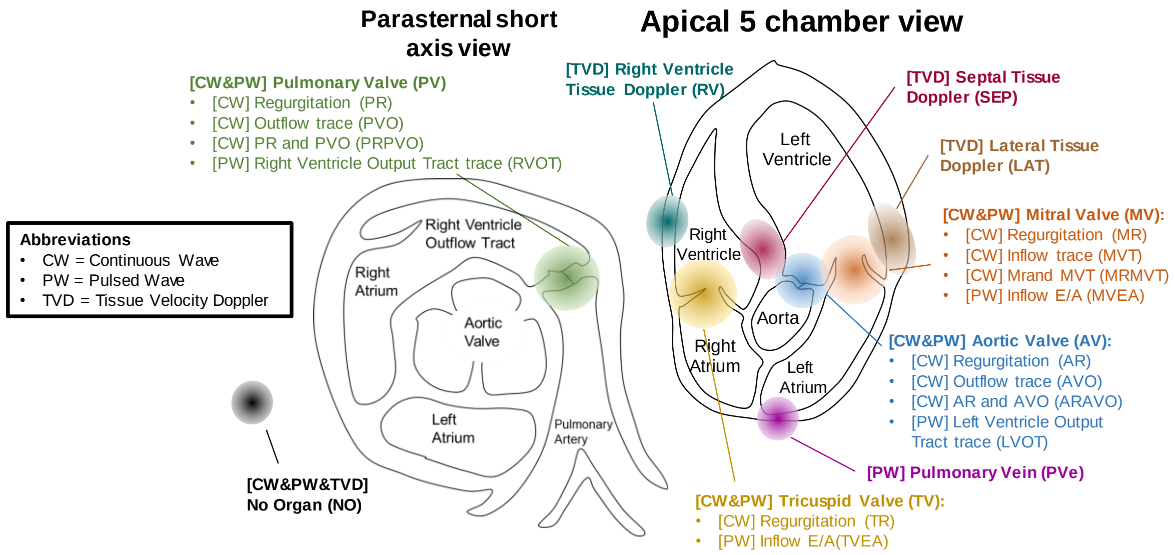
Parasternal short axis view
Apical 5 chamber view
[CW&PW] Pulmonary Valve (PV)
• [CW] Regurgitation (PR)
• [CW] Outflow trace (PVO)
• [CW] PR and PVO (PRPVO)
• [PW] Right Ventricle Output Tract trace (RVOT)
Abbreviations
• CW = Continuous Wave
• PW = Pulsed Wave
• TVD = Tissue Velocity Doppler
Right Ventricle Outflow Tract
Right Atrium
Aortic Valve
Left Atrium
Pulmonary Artery
[CW&PW&TVD] No Organ (NO)
[TVD] Right Ventricle Tissue Doppler (RV)
[TVD] Septal Tissue Doppler (SEP)
[TVD] Lateral Tissue Doppler (LAT)
Left Ventricle
Right Ventricle
Aorta
Right Atrium
Left Atrium
[CW&PW] Mitral Valve (MV):
• [CW] Regurgitation (MR)
• [CW] Inflow trace (MVT)
• [CW] Mrand MVT (MRMVT)
• [PW] Inflow E/A (MVEA)
[CW&PW] Aortic Valve (AV):
• [CW] Regurgitation (AR)
• [CW] Outflow trace (AVO)
• [CW] AR and AVO (ARAVO)
• [PW] Left Ventricle Output Tract trace (LVOT)
[PW] Pulmonary Vein (PVe)
[CW&PW] Tricuspid Valve (TV):
• [CW] Regurgitation (TR)
• [PW] Inflow E/A(TVEA)

Fig. 2. Each of the Doppler measurement types sorted by the location of the cursor position. Each color corresponds to a different region in anatomical space. The mode of each measurement in shown in front. Apical 5 chamber and parasternal short axis views are shown here for illustrative purposes only, to demonstrate the relative positions of the classes. Doppler spectra are typically acquired from a variety of echocardiographic views (see Appendix A for details) and part of the challenge of this problem is that the spatial relationship between structures demonstrated above will change depending on the view used for image acquisition. The No Organ (NO) class refers to images of air and ultrasound gel.

adjusted by the user during acquisition to focus on a specific part of the spectrum and prevent aliasing.

- The **mode** provides information on how the Doppler spectrum was acquired. Spectral Doppler incorporates three main imaging modes: Continuous Wave (CW) Doppler, Pulsed Wave (PW) Doppler, and Tissue Velocity Doppler (TVD). CW is used to measure high velocity blood flow across valves, PW provides flow analysis at specific spatial points, and TVD provides quantifiable myocardial velocities.
- The 2D **B-mode** (brightness mode) image shows the orientation of the probe with respect to the physical anatomy of the heart. Doppler spectra can be obtained from a variety of probe positions and angles depending on the desired measurement. The scan converted B-mode image is displayed here to orient the user.
- The **Doppler cursor,** visible on top of the B-mode image, indicates the spatial location of the spectrum. This parameter is interpreted in the context of the B-mode image. See Fig. 2 for a visual overview of how the cursor location corresponds to specific points in anatomical space. In the TVD classes the cursor is focused directly on the tissue, while in the CW and PW classes the cursor is focused on an area of blood flow. Exact positioning will depend on the desired measurement, operator preference, and individual patient anatomy.

Together, this information identifies the Doppler spectrum and therefore which measurements should be performed.

B. Clinical Need for Measurement Type Classification

Accurate automatic classification of Doppler measurement types can be combined with already available automated measurement techniques (e.g. [3], [4]) to provide fully automated analysis of Doppler spectra. Specifically, in a fully automated workflow, as soon as a Doppler exam is acquired the classification system is triggered and determines the measurement type. The system then triggers the corresponding automated measurement algorithm to display the measurement with no additional user interaction. This workflow is more efficient, allowing clinicians to spend more time on difficult measurements.

Furthermore, many clinics have petabytes of patient data in their archive systems from tracking patients over time. Thus, if used in combination with automated measurement techniques, one application of automatic Doppler measurement type classification is to perform rapid historical analysis on past exams in a robust and standardized manner. All information used in the proposed classification system is readily available in hospital archives if those archives store the raw data for each patient. Knowing a patient's progression from previous checkpoints can provide further information to support therapy planning. Therefore, historical analysis would provide clinical value through objective study of measurements over time. Another application is continually performing analysis on patients, which could bring statistical power to the development and augmentation of clinical guidelines.

C. Related Work

1) Ultrasound Classification

Doppler measurement type classification is unique because of the heterogeneity of data available in each classification example. As shown in Fig. 1, each recording contains image data, spectral data, modal parameters, a baseline position, and Doppler cursor coordinate locations. Previously, many of these items have been automatically classified individually, borrowing techniques from non-medical domains. Processing of spectral data has been a common task for several decades in speech recognition [5], and these techniques have been applied to Doppler spectra as well. For example, Wright *et al.* used artificial neural networks to classify Doppler spectra from arteries [6]. Meanwhile, automatic image classification has also become increasingly common as CNNs have surpassed the accuracy of humans on many tasks. Recently, these techniques were applied to echocardiographic B-mode images to automatically classify cardiac views with high accuracy [7], [8].

2) Multi-modal Learning

In non-medical fields, several groups have also looked at how data from different modalities can be combined. Ngiam *et al.* showed how a deep autoencoder could be trained with both video and speech data to generate a shared representation [9]. Ephrat *et al.* demonstrated how video and speech data could be encoded separately and then combined in a bidirectional long short-term memory network to solve the cocktail party problem of singling out a single speaker in a noisy audio track [10].

While many deep learning techniques have successfully made the transition from non-medical to medical applications [11], applying multimodal learning techniques remains a challenge because there are several orders of magnitude difference in the amount of available data. For example, Ephrat *et al.* were able to use >2000 hours of automatically annotated data. The annotation of such a volume of data in the context of Doppler spectra is challenging due to the lack of available simulated data. Transfer learning and fine tuning have previously been applied to solve data magnitude problems in medical imaging [12]. However, it is of limited use here since task objectives are different, and the relationship between the modalities (Doppler spectrum to B-mode) varies for each Doppler measurement class.

3) Confidence

One challenge in ultrasound imaging is that images acquired in clinical settings are not necessarily in standard views. During training, models are exposed to only a subset of possible views that might be seen in a clinical workflow. This is a concern in the given classification problem where misclassifications are more costly than doing nothing. Therefore, an algorithm to classify such images needs a mechanism to handle non-standard cases. This can be either collecting large datasets that can cover all possible views (even those that are non-standard) or a mechanism to bail-out when the image doesn't fall in the label set, such as via confidence metrics with a set threshold for acceptance.

Several groups have looked at how networks can give a confidence prediction along with an output label. It is well known that CNNs are prone to overfitting and cannot generalize well from the training set to unseen inputs [13]. Previously, Bayesian models have been used to provide a better estimate of model uncertainty by encoding model weights as a probability distribution. However, Bayesian techniques often come with increased parameter count and a higher computational cost to adequately model random distributions [14]. Monte Carlo dropout (MC-dropout) uses dropout at test time to approximate Bayesian inference with a lower computation cost [15]. Other methods such as temperature scaling [16] or histogram binning [17] calibrate fully trained network outputs without changing inference. Parameters are learned on the validation set to map network outputs to a true confidence distribution. These methods have the advantage of maintaining inference time and increasing the interpretability of the results without sacrificing the accuracy of the model.

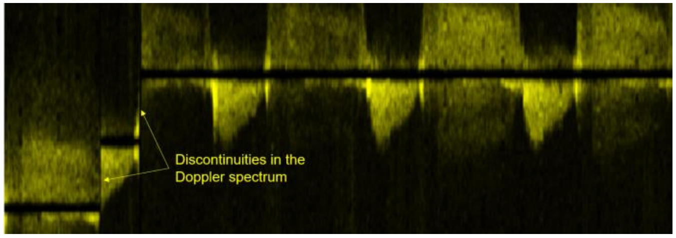

Fig. 3. Discontinuities arise when the operator shifts the baseline during acquisition. This is common practice when acquiring several measurements.

D. Contributions

After an analysis of the data, the spectral information was eliminated from the processing pipeline. This reduced the input to a B-mode image, Doppler cursor coordinate location, baseline position, and mode parameter. Although spectra provide useful information (and are used by clinical experts when labelling images), there are many variations in the collection of spectra that make it difficult to use in a network. For example, as shown in Fig. 3, spectral data can have discontinuities in the baseline as the user changes the parameters during acquisition. Spectral data is also variable length, which effectively shrinks or expands the features in the output image. Dealing with variable length would require an even larger dataset, since CNNs are not magnitude invariant. To avoid adding unnecessary complexity, the method developed here does not rely on spectral data. Instead, the method is focused on the integration of the latter four parameters. The spectra can be eliminated because we develop a novel pipeline which breaks the problem into a series of simpler pieces. We create an alternate way of uniquely identifying the spectra using these pieces. Our pipeline is outlined in Fig. 4 with references to the relevant section numbers for each piece. In brief, the principle contributions of this work are four-fold:

(1) **Heatmap encoding:** We show how to encode spatial features at the input of CNNs when multi-modal data includes coordinate locations as features.

(2) **Multi-head output:** We borrow techniques from multi-task learning to develop a multi-head learning strategy that integrates mode information to prevent misclassifications and reduce network size.

(3) **Decision tree mapping:** We use decision trees to incorporate user-defined imaging parameters in order to simplify the task of the CNN and better predict user intentions for the desired measurement type.

(4) **Confidence Thresholds:** We demonstrate how neural network layers, besides just the final layer, can be used to define a confidence metric that will disregard many images that differ from the training set. Our method requires no extra trained parameters, uses a fully nonlinear mapping between the output values and the network confidence estimate, and can be dynamically modified at inference time depending on the desired tradeoff between ignored and error rates.

To the best of our knowledge, this is the first work to use CNNs to classify Doppler measurement types. We achieve high

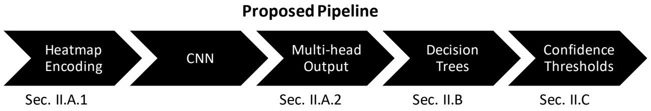

Fig. 4. The pipeline of the proposed method with relevant section numbers.

accuracy on the task, while maintaining a small memory footprint and close to real-time performance. Moreover, several of the methods developed in this work may be applicable to other classification problems, especially in medical imaging.

II. Methods

Through conversations with clinical experts, 18 of the most common measurement types for adults were identified. Three additional types of Doppler measurements were identified but were excluded from the current algorithm design due to infrequent clinical use. Specifically, among a dataset of over 7000 random Doppler images collected from a clinical site, just 30 images came from these three classes combined. Including these measurement types would be more likely to confuse the network and would require a significant effort to collect sufficient training data. Additionally, including the classes would not result in noticeable clinical time savings after algorithm implementation since they are used infrequently. Two steps are taken to account for measurement types not covered in our label set. First, to avoid making a classification on images scanned without a visible B-mode image on screen, a no organ (NO) class was added which consisted only of images where air and varying amounts of ultrasound gel were scanned. The Doppler cursor, baseline, and other parameters were chosen to cover a variety of possible inputs for the NO images. Second, a confidence metric was designed (Sec. II.C) to discard images from other classes. A full discussion of each of the measurement types is outside the scope of this work, but Fig. 2 shows a diagram of the relative cursor positions as well as the abbreviations for each type. An outline of each measurement's use and acquisition is available in [1], and reports specific to CW and PW [18], and TVD [19] mode measurements are also available. In addition, a further description of each measurement type is presented in Appendix A of the supplementary material.

The proposed method performs a classification on these Doppler measurement types. The relevant anatomical region is determined using a CNN as described in Sec. II.A. Sec. II.A.1 explains heatmap encoding at the input of the network while Sec. II.A.2 describes a multi-head output approach to divide the classification according to the imaging mode. A decision tree to simplify the network's classification task is presented in Sec. II.B. A confidence metric is defined in Sec. II.C to avoid misclassifications for low-confidence cases such as images from other measurement types or images with poor quality. Finally, the design of the dataset used for training and testing is outlined in Sec. II.D.

A. Determining Cursor Location with CNNs

As shown in Fig. 1, a single Doppler recording is composed of many multi-modal features. Given the information in the format of Fig. 1, an expert observer can mentally integrate the relevant information and classify the type of measurement that should be made. However, it would be unrealistic to expect a network to perform a classification given only an image such as Fig. 1 because some of the most important pieces of information are not emphasized in the image. For example, the Doppler cursor is very important to the classification because it indicates the location of the Doppler spectrum within the heart, but it is only a small marker on the image.

Instead, all the relevant data is extracted individually from each recording. The mode is recorded as either CW, PW, or TVD. The relative baseline is extracted as a float in the range from 0 to 1, where the default (unchanged) location is 0.5. The raw B-mode data is extracted as a 512×256 image, since the depth dimension is usually much larger in the raw data. Note that the non-scan converted (beam space) data is used directly rather than the scan-converted (probe space) data that is shown to the user. The added step in the pipeline to scan-convert the images yields no gain in this application where the Doppler cursor position relative to the heart structures is the key piece of information. Scan-converted images could equivalently be used.

As shown by the different colors in Fig. 2, the measurement types can be grouped into 9 locations in anatomical space. Since the relationship between cursor coordinates and image features would be similar for each of these locations, all measurements from the same anatomical location are merged into the same class for the CNN. Thus, the task of the CNN is only to figure out which anatomical location the measurement came from, the rest is handled during post-processing as described in Sec. II.B. The only inputs into the network are the B-mode image and the cursor coordinate.

1) Heatmap Encoding

The position of the cursor is extracted relative to the original B-mode image as a coordinate pair. In the proposed approach the coordinates are encoded as a heatmap. The coordinates are not directly used because Liu *et al.* showed CNN's are typically poor at learning mapping between coordinates in cartesian coordinate space and pixel space [20]. Additionally, in landmark detection problems, the current state of the art is to extract landmark coordinates from heatmaps of likely locations produced by the network [21]. Intuitively, using heatmaps works because there is a linear mapping from the coordinate space of the input image to the output heatmap. Logically, networks should also perform better if landmarks at the input are encoded as heatmaps instead of input as coordinates. To encode the Doppler cursor location as a heatmap, we generate a 2D normal gaussian probability density function with a standard deviation of 10 pixels centered at the cursor coordinate. The heatmap is generated in 512×256 resolution to match the original raw data, and then appended to the input image as an additional channel. Image and heatmap are both rescaled to 256×256, which has the effect of compressing the gaussian vertically. This allows the expected spatial distribution of the landmark to more closely match the physical dimensions of the raw data. An example heatmap is shown in Fig. 5. Finally, both image and heatmap are cropped to 224×224. During training, random crops are used for augmentation. Center crops are used during validation and testing.

2) Multi-Head Network

As shown in Fig. 2, other than Pulmonary Vein, the CW and PW mode measurements share the same anatomical locations.

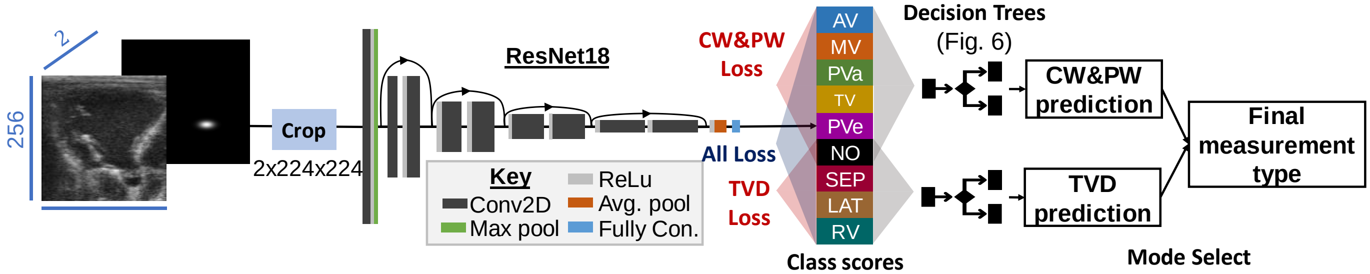

Fig. 5. The heatmap of the Doppler cursor location is appended as a channel to the non-scanconverted B-mode image and both are cropped to 224×224 before being input to the network. A ResNet18 [28] is presented here, but a variety of network architectures are tested (section III.A.3). For all networks, the last fully connected layer (of size 9) is split into two groups (heads). The No Organ (NO) class is input to both heads so the CW&PW head is 6 classes and the TVD node is 4 classes. During training the loss can be backpropagated from each head individually, or together from all classes (section II.A.2) During inference each head is passed to its own classifier and decision tree (see Fig. 6 for decision trees). The mode parameter is used to select between the two outputs to yield a final class. See Fig. 2 for class abbreviations.

The CW and PW locations are also completely distinct from the TVD mode measurements (except for the no-organ synthetic class). Because the mode parameter is always set by the user, simply training a network to classify all outputs would lead to unnecessary errors. The network should never classify a TVD mode image into a CW/PW measurement type since the mode is known at classification time, but without explicit separation this misclassification may occur. To solve this, the set of anatomical location classes can be split into unique sets, one for CW&PW and one for TVD. One approach would then be to train a different network for each mode and call the network for the relevant mode during inference. However, this approach doubles the memory footprint of any implementation, which is a downside for integration into a resource-constrained environment.

An alternative solution is to frame this as a multi-task learning problem. Multi-task learning integrates the information from several related tasks into a single network by implementing a separate network branch for each task. Often in multi-task networks, information from one task improves performance on another. The approach has proven to be successful in a variety of deep learning applications [22]. Our method is a slight variant of multi-task learning adapted for this problem. In detail, the network presented here has only a single branch with multiple classification heads operating on the final layer. Only one head is relevant for every given input sample. Therefore loss is backpropagated only from the classification head with a mode matching that sample whereas in multi-task learning loss can be taken from all branches during training. Input to each classification head are the values from the last fully connected layer for the classes that belong to that mode. The total loss for a given minibatch is shown in (1) where $f(x;\theta)$ is the output of a network with input x and parameters θ, x_{TVD} and x_{CWPW} are the TVD and CWPW samples in the minibatch respectively, and CE is cross-entropy loss. The λ values control the weighting between heads.

$$L = \lambda_1 CE(f(x_{TVD};\theta)) + \lambda_2 CE(f(x_{CWPW};\theta)) \quad (1)$$

During inference, the CNN yields predictions from both heads, but only the value from the relevant mode is read by the calling function. Due to these differences we instead call this a "multi-head" network. With this design choice, we exploit the information about the different modes by including separate heads and loss functions for the CW&PW and TVD groupings of anatomical locations. The architecture of the multi-head network is shown in Fig. 5.

B. *Decision Tree Mapping*

After finding the anatomical location with the CNN, the next step in the pipeline is determining the final measurement type. The mode and relative baseline position parameters extracted from the spectrum linearly separate the measurement types in each anatomical region because users change those parameters based on which type of measurement they wish to acquire. Therefore, decision trees are used for post-processing the output of each head as shown in Fig. 6. One possible error is introduced in this scheme when a CW image is classified as a Pulmonary Vein (PVe). In preliminary experiments this was never an issue, but occurrence in a clinical setting would require manual re-classification.

Decision trees are a better solution than feeding the parameters into the network because it avoids unnecessary mistakes and enables the CNN to use classes that are based solely on regions in anatomical space. Otherwise, the CNN would likely be confused between classes from the same anatomical region. Additionally, several of the original smaller classes do not have enough images for a network to properly converge. Grouping the classes increases performance.

C. *Confidence Metric*

Correct classifications from the network will yield significant time savings for clinical users by automatically launching the tool associated with that Doppler measurement type, where available. However, incorrect classification comes with a cost as the user will have to navigate back in the menu and select the correct measurement type. As automation continues to permeate clinical workflows, this cost may become larger. Initial misclassification could trigger unrelated measurements and automated tools. Moreover, there may be images in a clinical setting that are different from those seen during training. Thus, it is important for the network to have a bail-out mechanism on images with high uncertainty.

Our approach relies on the last fully connected layer before the softmax classifier, named the "pre-softmax" layer. This layer was chosen because raw network estimates for all classes are readily available before distortion by the multiple heads. The pre-softmax values for each example in the validation set are recorded after the network weights are trained and frozen. The recorded values are divided into quantiles. That is, rather than learning a mapping from outputs to true confidence (as was done in [16] and [17]), a series of cutoff values are found for each confidence level. During test time, the quantile is set based

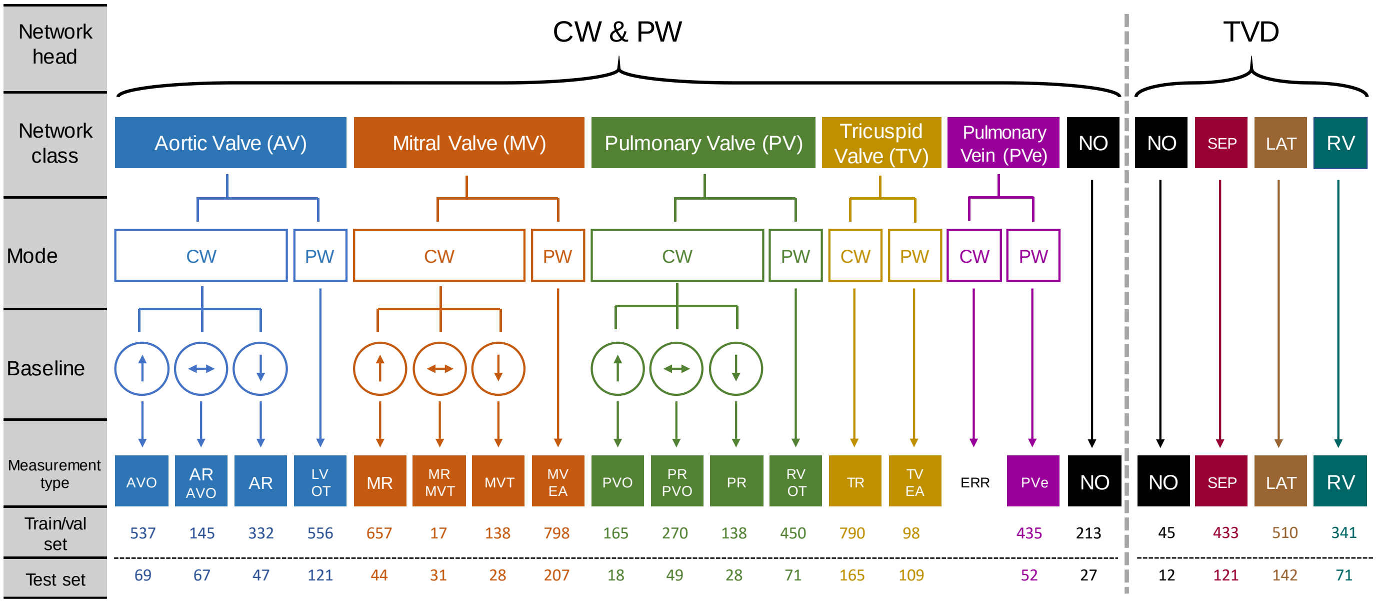

Fig. 6. Decision trees from the classes output by the network to the final classes determining the measurement type. The network has two heads, the CW&PW head and the TVD head. Outputs from each head are mapped to final classes using the mode and baseline parameters. The mode is then used to decide between the CW&PW head output and the TVD head output (Fig. 5). For the baseline: (↑) indicates a baseline was moved upwards – towards positive values, exposing a larger range of negative velocities, (↔) indicates a baseline is in center position, and (↓) indicates the baseline was moved downwards – towards negative values, exposing a larger range of positive velocities. If nothing is indicated for mode or baseline, then those parameters are not used for that mapping (all values map to the same class). For example. every image in the TVD head is guaranteed to be mode TVD so the mode is not relevant. Training, validation, and test set sizes are shown below each class. See Fig. 2 for locations of each class and acronym definition. ARAVO, MRMVT, and PRPVO are combinations of AR/AVO, MR/MVT, and PR/PVO respectively.

on the desired tradeoff between error rate and ignored rate. The maximum output value is found as usual, but if the pre-softmax value for that class falls below the given threshold then the image is labeled as low confidence and ignored.

To validate the chosen approach, results were compared to several other methods of determining confidence. MC-dropout has proven to be one method for approximating Bayesian inference in a computationally efficient manner [15]. An MC-dropout version of the model is implemented following the approach in [23], where 50% of the neurons from the last fully connected layer are randomly dropped during each inference run. In addition, combining the predictions of ensembles of neural networks has given superior classification performance [24], [25]. However, ensembled outputs can also be used as a measure of algorithm confidence. An ensembled confidence is implemented by ignoring cases where networks within the ensemble predict differing classes.

D. Dataset

The training and validation dataset consisted of exams previously collected by GE Healthcare for internal tool development. All exams were fully anonymized and came from a single clinical site. Exams were collected to try to maintain a high number in each class, but more images were available for classes that occur more frequently in clinical practice than those that occur infrequently. Thus, the dataset is slightly unbalanced because it reflects the clinical distribution of the data. Note that all classes of the same color in Fig. 2 are grouped together for training the network and split later in post-processing as shown in Fig. 6. The final set was 7081 images where individual class sizes are shown in Fig. 6.

Exams from seven institutions were used for the test set. The test institutions were spread over six different countries and three different continents. All test set institutions were different from the training set institution. This was done for two reasons. First, since images are fully anonymized, it is impossible to guarantee that two images from the same institution are not from the same patient. It is crucial for accurate test statistics that the training and test sets contain unique patients. Second, every institution has slightly different acquisition practices and patient populations, leading to small differences in the distribution of the images. Thus, to get a result that reflects real performance "in-the-wild" it is important to test on data from separate institutions. The test set contained 1479 images and class distributions are also shown in Fig. 6. All images were labeled by a clinical expert experienced in Doppler spectrum analysis and reviewed for accuracy by two other experts. Roles were swapped between sets, so a different expert did the initial labeling for the test set.

While gathering the training and validation sets, there were 298 images that had insufficient image quality for an expert to classify them. These images were categorized as the *unknown* set to analyze the confidence metric. Additionally, 30 images were identified that belonged to the three measurement classes not included in this network because they appear infrequently in clinical practice. These images were put into the *extra* set to analyze the confidence metric. Anonymization procedures removed all patient information, so the number of patients in the datasets is unknown.

E. Testing

To validate reproducibility, the combined training and validation set is used to estimate five different models. Each model is trained using 90% of the dataset with the remaining 10% set aside as validation. The model with the best

#	Architecture	Input	Networks	Classification Heads	Accuracy			F_1 Score (std)	Size (MB)	Time (ms)
					TVD	CWPW	Total (std)			
E1	ResNet18	I	2	One head per network	52.5%	70.8%	67.3% (0.4%)	63.1% (0.7%)	1480	3.5
E2	ResNet18	H	2	One head per network	83.9%	66.2%	68.8% (0.3%)	70.3% (1.3%)	1480	3.5
E3	ResNet18	I + H	2	One head per network	96.9%	95.3%	95.7% (0.3%)	96.3% (0.2%)	1480	3.5
E4	ResNet18	I + H	1	One head	90.8%	94.5%	93.7% (0.4%)	94.3% (0.3%)	**740**	3.5
E5	ResNet18	I + H	1	Training: one head Testing: two heads	98.4%	**95.7%**	**96.4%** (0.3%)	**97.1% (0.3%)**	**740**	3.5
E6	ResNet18	I + H	1	Two heads	**98.8%**	95.0%	95.8% (0.9%)	96.6% (0.7%)	**740**	3.5

Table 1. Comparison of experimental results for different input and output settings. In the Input column, I indicates only an image, H indicates only a heatmap, and I + H means the image with heatmap concatenated as shown in Fig. 5. In the Classification heads column "one head" refers to a single softmax classifier with all classes, and "two heads" refers to the multi-head approach detailed in Fig. 5. Accuracy is total correct images over total images (weighting classes with more images more) and F_1 score is an average across the individual F_1 scores of each class (weighting each class equally).

performance on each validation set during training is saved for testing on the independent test set. The train/validation/test split as a percentage of the total data is 74%/9%/17%. The five different validation sets are non-overlapping. Quantile cutoff limits for the confidence metric are extracted by averaging results across the five validation sets. This setup is used for all presented approaches and metrics are averaged from evaluating all five trained models on the unseen test set. Using five different models is important to (a) better estimate accuracies on the test set, (b) provide more robust quantile cutoff limits extracted across a larger set of unseen examples, and (c) obtain different models for the ensemble-based confidence method.

III. Results

To evaluate the effects of each design decision, a series of experiments were constructed with metrics measuring accuracy, speed, and memory usage. Classification accuracy was measured as defined in (2) where N_I is the total number of images in the test set, C is the set of classes, and TP_i is the number of true positives for class i. The F_1 score [26], was measured as a combination of recall and precision (3). The recall of class i, R_i, is given by $R_\mathrm{i} = TP_\mathrm{i}/(TP_\mathrm{i} + FN_\mathrm{i})$ where FN_i is the number of false negatives. The precision of class i, P_i, is given by $P_\mathrm{i} = TP_\mathrm{i}/(TP_\mathrm{i} + FP_\mathrm{i})$ where FP_i is the number of false positives. N_c is the number of classes.

$$Accuracy = {}^{1}\!/_{N_I} \sum\nolimits_{i\in C} TP_\mathrm{i} \tag{2}$$

$$\mathrm{F}_1 = \frac{2 * \sum_{i\in C} R_\mathrm{i} * \sum_{i\in C} P_\mathrm{i}}{N_C * (\sum_{i\in C} R_\mathrm{i} + \sum_{i\in C} P_\mathrm{i})} \tag{3}$$

Note that accuracy was obtained by micro-averaging (weighting by class frequency), while F_1 was obtained by macro-averaging (weighting each class equally). Although micro-averaging will bias results towards classes with more images, it also reflects the clinical reality since classes with more images will appear more in clinical use. The memory size and inference time measurements were implemented following [27]. We first examine the effect of different input and output settings in the experimental setup (Table 1) and then look at the performance of various network architectures (Table 2). For the first experiments a ResNet18 network (architecture shown in Fig. 5) was chosen because the residual connections in ResNet speed up training and improve accuracy when training deeper networks [28]. Specifically, ResNet18 has a smaller footprint than other networks and less parameters, which helps avoid overfitting on data-limited tasks.

A. *Cursor Location with CNNs*

1) *Heatmap Encoding*

First, the effect of adding the cursor heatmap was evaluated. A network was trained using only the B-mode image as an input (E1 in Table 1), using only the heatmap as an input (E2), and then with the cursor heatmap appended to the B-mode image (E3). In all three cases, separate networks were trained for each mode (CW&PW vs. TVD). As expected, with only a single input channel (either image or heatmap) there were low classification accuracies. The TVD network in E2 (heatmap only) did achieve 84% accuracy which shows the heatmap provides quite a bit of information for TVD cases. This is intuitive since TVD images are almost always acquired from the same echocardiography view (apical four chamber). Therefore, the position of the cursor remains relatively constant within each class and different between them. Conversely, results were worse with only a heatmap for the CW/PW mode. In this mode views (and therefore cursor locations) change within classes. Results showed a significant improvement with both the image and heatmap passed to the network (95.7% accuracy).

2) *Multi-head Networks*

Second, the effect of the multi-head approach was tested. As a baseline approach one network was trained with a single classification head on all 9 classes (E4). There are 9 classes instead of 10 here because with a single classification head only one NO class is needed. Results showed a 2% drop in accuracy compared to E3, indicating that not splitting the classes creates a harder task for the network. However, the memory footprint was also cut in half. To test whether the multi-head approach could achieve the same accuracy as two separate networks (E3) with the footprint of a single network, the multi-head architecture was applied at test time to the network trained in E4 (E5 in Table 1). Next, a single network was trained using the multi-head approach during both training and testing (E6). Experiments E3 – E6 all used the same input information, but with different methods of integrating the mode information.

Results showed that using a single head at training, but multiple heads at test time (E5) resulted in the best performance

#	Architecture	Accuracy	F_1 score	Size (MB)	Time (ms)
E5	ResNet18 [28]	96.4%	97.1%	740	**3.5**
E7	ResNet34 [28]	96.5%	97.1%	911	6.2
E8	ResNet50 [28]	96.2%	97.0%	966	8.8
E9	SqueezeNet-v1.1 [31]	94.2%	94.8%	584	3.9
E10	ShuffleNet [30]	96.0%	96.8%	**579**	12.0
E11	MobileNet-v2 [32]	96.0%	96.8%	582	7.4
E12	NASNet-A-Mobile [29]	**96.7%**	**97.3%**	653	46.3
E13	GoogLeNet [25]	96.1%	96.8%	702	10.0
E14	DenseNet-121 [33]	96.5%	**97.3%**	653	22
E15	DenseNet-169 [33]	96.2%	97.0%	735	30.9
E16	DenseNet-201 [33]	96.5%	**97.3%**	828	36.8
E17	BN-Inception [34]	96.2%	96.8%	697	13.9
E18	DualPathNet-68 [35]	96.4%	97.2%	735	27.7
E19	Xception [36]	96.4%	97.2%	891	8.0
E20	Inception-v3 [37]	96.0%	96.7%	908	15.8

Table 2. Comparison of experimental results with different network architectures. Accuracy and F1 scores are means across five networks trained using the setup described in section II.E. All experiments used the same input/output configuration as E5 in Table 1.

(96.3% accuracy), slightly better than using separate networks (E3). This result indicates that the network can use the information from other modes to improve performance like multi-task learning. It was also better to use the multi-head architecture only during testing (E5) than both training and testing (E6), possibly because the harder task of training with all 9 classes forced the networks to better differentiate between classes.

3) Network Architecture

Third, the impact of network architecture was tested. The top 15 architectures in terms of Top-1 accuracy density from the benchmark analysis of Bianco et al. [27] were evaluated. Top-1 accuracy density is classification accuracy (in their work on ImageNet) divided by number of parameters and is a measure of a network's performance relative to its size. Details on the implementation of all networks are provided in Appendix B. The E5 configuration was used for all architecture experiments.

Results are presented in Table 2. The NasNet-A-Mobile [29] (E12) had the best accuracy (96.7%), while ShuffleNet [30] (E10) had the smallest size (579 MB), and ResNet18 (E5) had the fastest inference time (3.5 ms). All models achieved accuracies >94% indicating that our method is resilient to changes in architecture. The chosen architecture will depend on the desired tradeoff between accuracy, memory size, and speed, but we judged the optimal approach considering all factors to be the ResNet18 architecture (E5). ResNet18 has just slightly lower accuracy (96.3%) than NasNet-A-Mobile but is more than 10 times as fast.

The remainder of the results are presented using the single network from E5 with the best performance on its validation set. Using validation sets is not a fair comparison since every network has a different validation set. However, as shown in Table 1, the variance between results is quite low so the choice of network does not significantly affect the results. The confusion matrix on the test set for this network is shown in Fig. 7. Tricuspid Valve (TV) had the lowest accuracy with 93%. Despite the uneven distribution of training data between classes, the method does not overfit to the classes with more images.

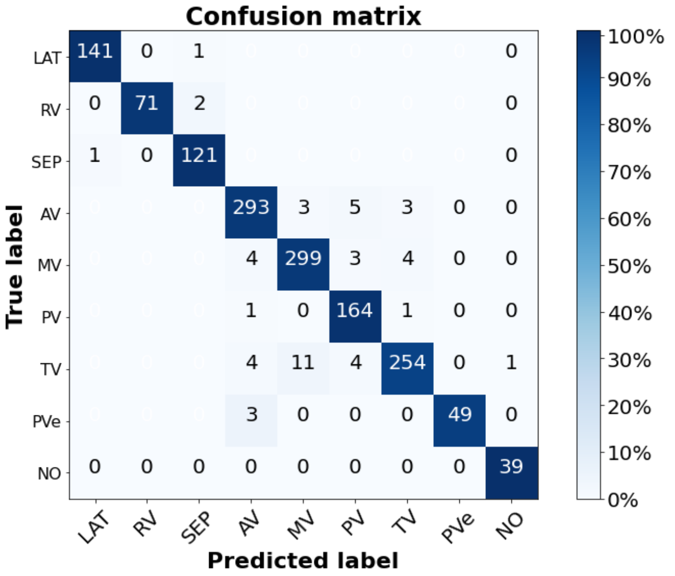

Fig. 7. Confusion matrix on the final test set using E5 The boxes where no number is shown will never be misclassified because separate classification heads are used. Colors are normalized to class size (percentages). Abbreviations are described in Fig. 2.

B. Decision Tree Mapping

Using the decision tree presented in Fig. 6 the output of the network was mapped to the final measurement type classes. The accuracy for each type is shown in Table 3. To check what kind of mistakes were occurring, an error analysis was performed for the two measurement types with the lowest accuracy: Aortic Regurgitation (AR) and Tricuspid Regurgitation (TR). For AR, there were 4 misclassified images. Of these, 3 were acquired from the apical parasternal long axis view (A full description of echocardiographic views is available in [1]). It is logical the network might miss these cases since there were few AR images acquired from this view in the training set because AR measurements are typically taken from the apical 5 chamber view. However, operators at different clinics may have different preferences, leading to the discrepancy between training and test sets. The last image was judged to have been misclassified by the labeler on re-analysis.

There were 16 misclassified TR images. Of these misclassifications 13 images were from the right ventricle inflow view. This view was not included in the training set because the labeler for the training set did not have experience with this view and thus ignored those images. Therefore, the network never learned the patterns associated with these images. For one of the remaining images the class could not be determined during re-analysis, it was initially classified as TR because TR measurements were encoded in the file. The

TVD

SEP	LAT	RV
99.1%	99.3%	97.3%

PW

LVOT	MVEA	RVOT	TVEA	PVe
97.5%	96.1%	97.2%	96.3%	94.2%

CW

AVO	ARAVO	AR	MR	MRMVT	MVT	PVO	PRPVO	PR	TR
94.2%	98.5%	93.6%	95.5%	100%	96.4%	100%	100%	100%	90.3%

Table 3. Classification accuracy for the selected network for each measurement type sorted by mode. Abbreviations are described in Fig. 2.

remaining 2 images were simply missed by the network.

C. Confidence Metric

To test the validity of the proposed confidence metric, the pre-softmax set of cutoff values were extracted. Quantile cutoff limits were extracted for each class from 0%-10% quantiles in 0.2% step sizes. The quantile is the ignored percentage on the validation set: a quantile of 5% indicates that the 5% of images with the lowest pre-softmax values would be ignored. Quantiles were tested only on the single network but were obtained by averaging across all five networks and validation sets to enhance the robustness.

The test set was split into two pieces, the first containing the 1431 correctly classified images (*correct* set) from E5 and the second containing the 51 misclassifications (*incorrect* set). The quantiles were used when running inference on these two sets of images as well as the full *test* set and the *unknown* and *extra* sets put aside during labeling. For each set, the ignored rate was recorded while iterating through the quantile values. For the test set, the error rate was also recorded. Results are shown in Fig. 8, with ignored rates on the left axis and error rate (in red) on the right axis.

The confidence metric results may indicate some overfitting, as well as that the validation and test sets came from different distributions. If the network is not overfit and the images are from the same distribution, the quantile should map 1:1 to the ignored test percentage. However, results showed at the 0.5% quantile 5% of the test images are ignored. A difference in distribution is expected since the images came from separate clinics and matches what was found in the error analysis.

Results also demonstrated the confidence metric accurately detected which images came from outside our training distribution and eliminated misclassified images at a much higher rate than correctly classified ones. For example, a user setting the confidence threshold at the 8% quantile mark would have to manually label 16.5% of their images but would achieve an accuracy of 99% on those images automatically classified because ~80% of the *incorrect* images would be ignored.

Both the *unknown* and *extra* image sets were also ignored at a much higher rate, confirming that the metric identified out of distribution samples. Ignored rates showed an approximately logarithmic relationship with increasing quantile values for all three of the *incorrect*, *unknown*, and *extra* sets. For all three, >20% of images were ignored at the 1% quantile.

Results were compared to the MC-dropout versions of the model. Each MC-dropout model was run 100 times, and quantile limits were set for values from the pre-softmax layer and softmax layer of the normal model, and from the mean and variance of the pre-softmax and softmax layers from the MC-dropout model. Results were similar for all implementations, with a slightly higher ignored rate for all sets when using the pre-softmax layer from either model. These results indicate that the choice of how to extract quantile values does not play a large role in the resulting confidence metric. The advantages of using the pre-softmax values are that it is simpler to implement and no additional experiments need to be run.

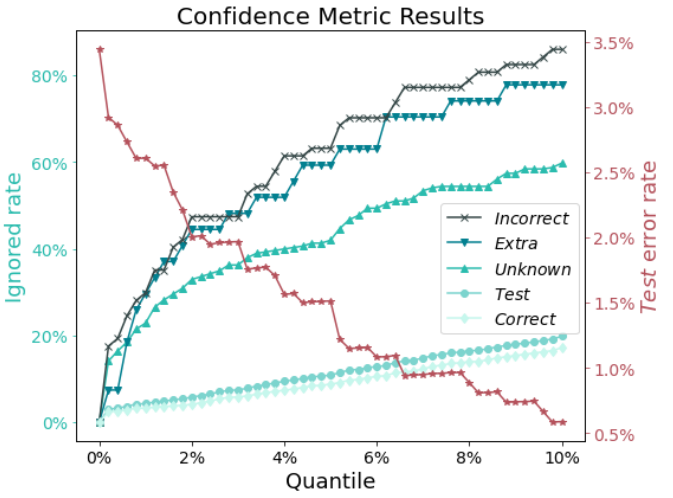

Fig. 8. Results of confidence metric experiments with ignored rates (green lines) on the left axis and error rate of the test set (red line) on the right axis. Ignored rate refers to the percentage of images which the network did not label because the output was beneath the cutoff for that quantile. *Test* refers to the test set of images. *Correct* is the subset of test images correctly classified by the network in E5 and *incorrect* is the subset of misclassifications. *Unknown* is the set of images which were unidentified during labeling. *Extra* is the set of images from classes not included in this classification scheme. An ideal confidence metric should completely ignore the *incorrect* and *extra* sets while ignoring none of the *correct* set. Ignoring the *unknown* set indicates the network accurately reflects expert confidence, but it is possible the network has detected features unseen to the observer. Best viewed in color.

Metrics from the ensembled approaches were also compared to selected quantiles of the proposed confidence metric in experiments C1-C5 in Table 4. The ensemble networks contained all five originally trained models. The predicted class came from either the majority vote (C4) or only images where all networks agreed while other images were ignored (C5). These ensembled approaches were compared to representative quantiles from Fig. 8: the 0% quantile (C1 – the single chosen network from E5), the 5.2% quantile (C2 – selected as the closest error rate to C5), and the 8% quantile (C3 – first quantile with error <1%). Using majority voting (C4) provided a small improvement in error rate (3.2%) compared to C1 (3.4%) without ignoring any images. Selecting only matching outputs (C5) provides a significant decrease in error rate (1.3%) although 4.8% of images were ignored. To achieve an equivalent error rate with the proposed confidence metric (C2), more than twice the number of images were ignored. The downside of the ensembled approach is five networks must be initialized, significantly increasing the memory consumption.

D. Implementation Details

All parameters were extracted automatically from private Dicom tags using Python 3.6. Pre-processing, training, and testing were carried out on an Ubuntu machine with Python 3.6, PyTorch 0.4, and an NVIDIA Titan X GPU. All networks were trained for 60 epochs. The learning rate was set to 0.1 for all experiments other than E9 where it had to be reduced to 0.01 to get the network to converge. Learning rates were reduced by a factor of 10 every 20 epochs. All experiments used standard gradient descent with momentum (coefficient 0.9) and weight decay (0.0005). We used normalization to center the dataset during training and inference. The B-mode images were normalized to [0,1]. The heatmaps were already in the range [0,1] because they are probability maps., Both images and heatmaps were mean-centered by subtracting the mean value of all pixels in all images in the training set. This value was 0.3 for the B-mode image and 0.0062 for the heatmap image. Cross-entropy loss was used for all experiments and the classes were not weighted since the results did not suffer from class imbalance. Preliminary experiments with a weighted loss function did not improve results. While training the multi-head network, λ_1 and λ_2 (loss function hyperparameters from (1)) were set to 0.18 and 0.82 respectively, but results were not sensitive to changes of λ's within normal ranges. TVD classification is an easier task (as shown in Table 1) so $\lambda_2 > \lambda_1$ even though there were fewer TVD images.

IV. Discussion

Our results indicate that highly accurate Doppler spectrum measurement type classification is possible in echocardiography *without* using the spectrum data. Accurate classification despite the omission of the Doppler spectra proves the network is learning the relationship between user input and anatomical structures. The spectrum data can be ignored because each class correlates to a unique physical location within the heart after using our mapping scheme. Note that since each Doppler spectrum can be acquired from a variety of different views, this does not imply each class corresponds to a unique location in the input image. Accurate classification requires understanding of both the B-mode image and the cursor location. Highly accurate results have already been achieved on view recognition tasks in echocardiography (e.g. [7], [8]) indicating effective understanding of the B-mode image through CNNs. Our results take this a step further. We show heatmaps are an effective way to encode physical location information for CNNs, demonstrating the ability to connect anatomical structural information (B-mode image) to relevant user input (Doppler cursor location) in a classification.

Since deep-learning algorithms deployed in clinical settings must frequently compete for resources, methods for decreasing resource utilization were analyzed. Results demonstrated that a multi-head classification could reduce the memory footprint when the classification task can be split into separate problems by external parameters. The multi-head networks (E5/E6) maintained similar accuracy levels to those of separate networks (E3) and higher accuracy than a single network trained with all classes (E4). The final implementation achieved sub-4*ms* inference time, indicating near real-time performance.

Our approach demonstrated high accuracies across varying network architectures. Additionally, training several networks from differing dataset splits showed consistent results with low variances in F_1 scores and accuracies. The repeatability of our results across architectures and data splits is another strength of the contributions.

#	Architecture	Error rate	Ignored rate	Size (MB)	Time (ms)
C1	Single ResNet18, 0% quantile	3.4%	0%	**740**	**3.5**
C2	Single ResNet18, 5.2% quantile	1.2%	11.4%	**740**	**3.5**
C3	Single ResNet18, 8% quantile	0.9%	16.5%	**740**	**3.5**
C4	Ensemble ResNet18, Majority vote	3.2%	0%	3700	**3.5**
C5	Ensemble ResNet18, Matching output	1.3%	4.8%	3700	**3.5**

Table 4. Comparison of selected quantiles to ensemble methods using 5 networks. Size and time estimates for ensemble approaches assume that inference can be run for all networks simultaneously which depends on the implementation.

We also conducted an error analysis of the mistakes made by the network. Encouragingly, the error analysis showed the network is accurately learning the image and heatmap patterns included during training. The errors seen were mostly due to differences between the training and test sets in echo views or mismatch in labeling practices. Because the network accurately learned the patterns it was exposed to, accuracy could continually be improved by gathering additional training data that covers misclassified cases.

Misclassifications can be costly in a medical setting. They can lead to confusion when analyzing patient data and mistrust in artificial intelligence-based tools. To attempt to reduce misclassifications, several measures were taken. First, a No Organ (NO) class was included in the training dataset to avoid classifying images of air and gel into another class. Second, cutoff limits were set based on output values from the last fully connected ("pre-softmax") layer for each class. Images with a score below the cutoff were ignored rather than classified. Overall, results indicated that the proposed confidence metric can significantly reduce the error rate by ignoring missed images in the test set at a much higher rate than correctly classified ones. The confidence metric also ignores the *unknown* and *extra* image sets at an approximately three times higher rate than those from the test set. Our method demonstrates one way to handle inputs from unseen distributions in a classification problem. Moreover, it allows a user to easily set the quantile limit depending on the desired tradeoff between the error and ignored rates.

Results testing ensemble networks showed these methods ignore fewer images for the same error rate compared to quantile cutoffs. Ensemble methods are a more robust confidence predictor for environments without resource constraints. Moreover, ensemble methods could easily be combined with the quantile cutoff approach discussed above to provide a robust, tunable ignored vs. error rate tradeoff.

In future work, we hope to extend this method to public Dicom data (and thus multi-vendor). This is much easier because we don't use the Doppler spectra in our pipeline; the B-mode image, Doppler cursor location, mode, and baseline are the only data needed. Although we extracted this information from raw data in the present work, all of it is also available from public Dicom tools. The information could thus be extracted from any vendor's Doppler data. The main difference in using publicly available data (and thus scan-converted images) is some overlay on the public B-mode image, potentially including color-flow. However, with a little effort we anticipate the ability to overcome this and make our method fully multi-vendor.

V. Conclusion

In this work, we demonstrated a CNN-based method for the automated classification of Doppler measurement types. An example integration within a clinical workflow is presented in Appendix C of the supplementary material. Notable performance gains were shown on the task by encoding the Doppler cursor as a heatmap and introducing a post-processing mapping scheme to simplify the problem. These methods would also be applicable to other tasks including location information as an input parameter and/or with linearly separable classes. We design a confidence metric capable of discarding a large proportion of images with high uncertainty to create a more reliable classification system. Our method performs fast and accurate classification of Doppler measurement types. In the same way automatic echocardiographic view recognition unlocked fully automated processing of many B-mode images, our work unlocks fully automated Doppler spectra analysis, bringing increased efficiency and statistical power to clinical workflows.

Acknowledgement

The authors thank Pablo Lamata and Julia Schnabel for discussions and valuable feedback on the methods.

Supplementary Materials

Supplementary materials are available here: https://adgilbert.github.io/cardiac-doppler-type-classification/

References

[1] C. Mitchell *et al.*, "Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography," *J. Am. Soc. Echocardiogr.*, vol. 32, no. 1, pp. 1–64, 2019.

[2] A. Alfirevic, "Back to the Future—Importance of Spectral Doppler," *J. Cardiothorac. Vasc. Anesth.*, vol. 33, no. 5, pp. 1467–1470, 2018.

[3] L. Mo, D. Becker, K. McCann, M. Honda, and S. Ishiguro, "Method and Apparatus for Automatic Tracing of Doppler Time-Velocity Waveform Envelope," 5,935,074, 1999.

[4] U. Lempertz and E. Sokulin, "Cardiac Auto Doppler White Paper," 2017.

[5] L. Deng *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc.*, 2013, pp. 8604–8608.

[6] I. A. Wright, N. A. J. Gough, F. Rakebrandts, M. Wahabs, and J. P. Woodcock, "Neural network analysis of blood flow Doppler ultrasound signals: a pilot study," *Eur. J. Ultrasound*, vol. 6, no. 5, p. S11, 2004.

[7] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–8, Dec. 2018.

[8] J. Zhang *et al.*, "Fully Automated Echocardiogram Interpretation in Clinical Practice," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," pp. 1–9, 2011.

[10] A. Ephrat *et al.*, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *arXiv:1804.03619*, 2018.

[11] G. Litjens *et al.*, "A Survey on Deep Learning in Medical Image Analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[12] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[13] S. Falkner, A. Klein, and F. Hutter, "BOHB: Robust and Efficient Hyperparameter Optimization at Scale," *arXiv:1807.01774*, 2018.

[14] Yarin Gal, "Uncertainty in Deep Learning," 2016.

[15] Y. Gal and G. Zoubin, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *arXiv:1506.02142*, 2015.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. of the 34th Int.Conf. on Machine Learning*, 2017, pp. 1321–1330.

[17] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. of the 8th Int. Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 694–299.

[18] N. B. Schiller *et al.*, "Recommendations for Quantitation of the Left Ventricle by Two-Dimensional Echocardiography," *J. Am. Soc. Echocardiogr.*, vol. 2, no. 5, pp. 358–367, 1989.

[19] K. K. Kadappu and L. Thomas, "Tissue doppler imaging in echocardiography: Value and limitations," *Hear. Lung Circ.*, vol. 24, no. 3, pp. 224–233, 2015.

[20] R. Liu *et al.*, "An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution," in *Advances in Neural Inf. Proc. Sys.*, 2018.

[21] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "Numerical Coordinate Regression with Convolutional Neural Networks," *arXiv:1801.07372*, 2018.

[22] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv:1706.05098*, 2017.

[23] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in *Advances in Neural Inf. Proc. Sys.*, 2017.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[25] C. Szegedy *et al.*, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, vol. 07-12-June.

[26] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

[27] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark Analysis of Representative Deep Neural Network Architectures," *IEEE Access*, vol. 6, 2018.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[29] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, "Learning Transferable Architectures for Scalable Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8697–8710.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size.," *arXiv:1602.07360*, 2016.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger,

"Densely connected convolutional networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 2261–2269.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015, vol. 1, pp. 448–456.

[35] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-Decem, pp. 4468–4476.

[36] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions."

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 2818–2826.