Pseudo-D: Informing Multi-View Uncertainty Estimation with Calibrated Neural Training Dynamics

Ang Nan Gu$^{[0000-0001-8926-2397]1}$, Michael Tsang[2], Hooman Vaseli[1], Purang Abolmaesumi $^{\star 1}$, and Teresa Tsang $^{\star 2}$

[1] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
guangnan@ece.ubc.ca

[2] Division of Cardiology, Vancouver General Hospital, Vancouver, Canada

Abstract. Computer-aided diagnosis systems must make critical decisions from medical images that are often noisy, ambiguous, or conflicting, yet today's models are trained on overly simplistic labels that ignore diagnostic uncertainty. One-hot labels erase inter-rater variability and force models to make overconfident predictions, especially when faced with incomplete or artifact-laden inputs. We address this gap by introducing a novel framework that brings uncertainty back into the label space. Our method leverages neural network training dynamics (NNTD) to assess the inherent difficulty of each training sample. By aggregating and calibrating model predictions during training, we generate uncertainty-aware pseudo-labels that reflect the ambiguity encountered during learning. This label augmentation approach is architecture-agnostic and can be applied to any supervised learning pipeline to enhance uncertainty estimation and robustness. We validate our approach on a challenging echocardiography classification benchmark, demonstrating superior performance over specialized baselines in calibration, selective classification, and multi-view fusion.

Keywords: Multi-View Learning · Echocardiography · Uncertainty · Training Dynamics

1 Introduction

Medical image-based diagnosis faces two key challenges. First, it is safety-critical, where diagnostic errors can have serious consequences. Second, image acquisition is inherently imperfect. For example, ultrasound imaging depends heavily on the sonographer's skill and patient-specific factors, making it difficult to capture images at the optimal angle consistently. These limitations can lower image quality and, in turn, compromise diagnostic accuracy. As a result, evaluating these systems requires more than measuring classification accuracy; it must also account

* P. Abolmaesumi and T. Tsang are joint senior authors.

for the model's uncertainty estimates. In challenging cases or when image quality is poor, the system should know to abstain from making a prediction and instead refer the case to a human expert. This capability is assessed through the task of selective classification. Moreover, arriving at a diagnosis often requires integrating information from multiple sources, such as different imaging modalities or several views of the same anatomy. Crucially, the fusion process must account for the varying degrees of uncertainty associated with each source. This capability is evaluated through the task of multi-view fusion.

We propose a method to enhance selective classification and multi-view fusion by improving uncertainty estimation (UE). We focus on aleatoric uncertainty (AU), the irreducible uncertainty caused by incomplete or ambiguous input data. Although AU is less common in traditional vision tasks, it is critical in medical imaging, where anatomical features can be obscured or ambiguous because of patient variability, the imaging process, and modality-specific factors.

A core challenge in estimating AU is the lack of ground-truth labels that faithfully capture uncertainty. Classification datasets typically provide a single, definitive label, even though expert assessments often reflect borderline cases or ambiguity. Moreover, the common use of "one-hot" labels fails to convey the nuanced reasoning needed in uncertain cases. To achieve this, the model's confidence should be better aligned with task difficulty.

We leverage Neural Network Training Dynamics (NNTD) to generate pseudo-labels that quantify uncertainty based on how confidently and consistently the model learns each sample during training. Rather than relying on a fixed label, we track the model's evolving predictions across epochs and treat this trajectory as a measure of sample difficulty. NNTD-based methods have proven effective in detecting label noise [23,26], improving classification [24], and producing more reliable uncertainty estimates [9].

We propose *Pseudo-D*, a novel technique which combines NNTD information from both the training and validation sets to created pseudo-labels which are calibrated at the sub-class level. This is particularly useful when certain sub-classes are harder to distinguish than others. We evaluate Pseudo-D on a challenging multi-view ultrasound dataset for aortic stenosis (AS) classification, which requires integrating information from multiple scanning planes and handling patient-specific variability in image quality. We demonstrate that training with *Pseudo-D* improves uncertainty estimation in standard deep learning classifiers, and outperforms specialized methods on selective classification and multi-view fusion tasks. Compared to existing approaches, our method better aligns model uncertainty with input-specific factors like image quality and anatomical visibility. Furthermore, *Pseudo-D* is agnostic to model architecture and requires minimal changes to integrate to existing customized workflows.

2 Related Works

Selective Classification. The task of selective classification (SC), or prediction with a "reject" option, was initially extended to deep learning by Geifman et

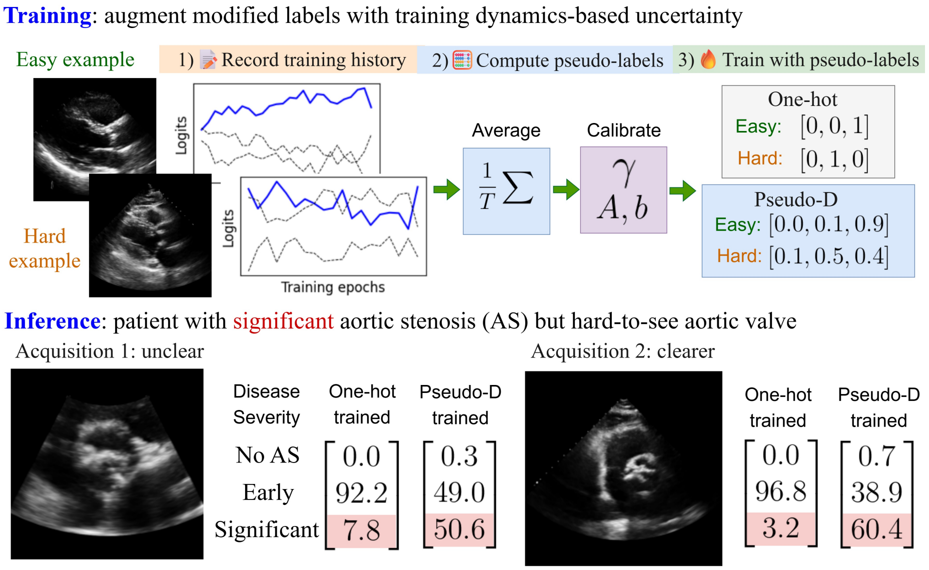

Fig. 1. We augment the training phase by first recording the history of predicted logits. The magnitude of the correct class logit (shown in blue) relative to other classes (in black) varies with task difficulty. We use the training history to generate pseudo-labels that align with the difficulty of each example. Our proposed training technique, *Pseudo-D*, yields a model with predicted probabilities that strongly correspond to image quality. Additionally, the pseudo-labels help mitigate overfitting by assigning lower confidence values to difficult training examples.

al. [8]. DeVries and Taylor [6] proposed explicitly learning uncertainty as an additional output of the model, using a modified loss function. Rabanser et al. [24] suggested using model checkpoints from different training epochs to form an ensemble for SC. Huang et al. [14] explored how label augmentation can improve SC, and Feng et al. [7] showed that the softmax response can outperform specialized scoring functions in existing SC approaches.

Multi-view Fusion. Multi-view fusion involves the combination of classifier predictions, where each prediction stems from a unique view of the same underlying object. The fusion of predicted probabilities can be through averaging [25], majority voting [19], or by learned weighting scheme [31]. Zhang et al. [33] establishes a theoretical link between uncertainty estimation and multi-view classification performance for logit-based approaches. Evidential Neural Networks (ENNs) [27] are trained to output belief masses instead of logits. Belief masses can be aggregated using Dempster's Rule of Combination, a mathematically rigorous method for combining multiple predictions [12].

Aortic Stenosis Severity Classification. Aortic stenosis (AS) is a heart valve disease characterized by restricted blood flow through the aortic valve. The clinical standard for AS diagnosis relies on measuring blood flow volume through the left ventricular outflow tract, typically derived from spectral Doppler [21]. However, Doppler-based diagnosis is sensitive to measurement variability [28,22] and is often unavailable on newer, lightweight ultrasound devices [10]. Recent clinical works [1,20] proposed assessing AS severity through B-mode ultrasound interpretation by human clinicians. AI-based AS classification using B-mode ultrasound has gained traction: works [5,11] classify using single-image or video inputs, while [2,13,15,17,30] and [32] utilize multiple views from retrospective exams, combining predictions via majority voting or averaging of predicted confidence. Huang et al. [16] uses multiple-instance learning to learn the importance of each image.

3 Methodology

3.1 Background: Construction of pseudo-labels via NNTD

Training samples $\mathcal{X}_{\text{train}} = (x_0, y_0), \ldots, (x_N, y_N)$ vary in how difficult they are for a model to learn. This difficulty may stem from intrinsic task complexity (e.g., distinguishing visually similar classes), data-related issues (such as noise, occlusions, or imaging artifacts), or label noise. The model's evolving predictions $f(x_i)$ over T training epochs provide insight into the difficulty of each data point. For easy examples, the logit corresponding to the ground truth class tends to dominate consistently across epochs. In contrast, for difficult or mislabeled examples, the logits for different classes often remain uncertain or competitive. We can compute pseudo-labels $y'_i \in [0,1]^C$ which represent the model's average class confidence over time:

$$y'_i = \frac{1}{T} \sum_{t=1}^{T} \sigma(f_t(x_i)), \qquad x_i \in \mathcal{X}_{\text{train}}, \tag{1}$$

where $f_t(x_i)$ denotes the model logits at epoch t, and $\sigma(\cdot)$ is the softmax function applied over class outputs.

These pseudo-labels encode the uncertainty related to sample difficulty, enabling more robust learning. Prior work has shown that training on such soft targets improves resistance to label noise [4,14] and enhances uncertainty estimation [9]. Training on y' can be viewed as a form of knowledge distillation from a training-dynamics-based ensemble, capturing model behavior across epochs. Crucially, these pseudo-labels are architecture-agnostic and can be used to train any downstream model on the same task.

3.2 Calibrated pseudo-labels

A limitation of Eqn. 1 is that the confidence of y'_i may not be well-calibrated, i.e. even if $f(.)$ fits these pseudo-labels, the predicted probabilities might still be over-

or under-confident. To address this, we generate pseudo-labels using information from both $\mathcal{X}_{train}$ and $\mathcal{X}_{val}$. Specifically, we apply temperature scaling to the logits so that output confidences better reflect the true accuracy (Eqn. 2 - 4):

$$v_i = \frac{1}{T}\sum_{t=1}^{T} f_t(x_i), \tag{2}$$

$$y_i' = \sigma(\gamma^*\ v_i), \qquad x_i \in \mathcal{X}_{\text{train}}, \tag{3}$$

$$\gamma^* = \underset{\gamma}{\text{argmin}}\ \text{CrsEnt}(\gamma\ v_j, y_j), \qquad x_j \in \mathcal{X}_{\text{val}}, \tag{4}$$

where $v_i \in \mathbb{R}^C$ denotes the logit vector averaged over epochs for both $\mathcal{X}_{train}$ and $\mathcal{X}_{val}$. The temperature parameter, denoted as γ^*, is chosen to minimize the negative log-likelihood (NLL), which is equivalent to minimizing cross entropy, on $\mathcal{X}_{val}$. We refer to this approach as *Pseudo-T*.

While temperature scaling is a common technique for improving model calibration, it typically applies a single global scaling factor across all classes. This uniform treatment fails to account for class-specific variability, since some classes can be inherently harder to differentiate compared to others.

To address this limitation, we adopt Dirichlet Calibration [18], which applies class-wise scaling to the model logits. This approach learns a transformation matrix and bias that adjust each class individually, yielding pseudo-labels with improved sub-class calibration:

$$y_i' = \sigma(A^* v_i + b^*), \qquad x_i \in \mathcal{X}_{\text{train}}, \tag{5}$$

$$A^*, b^* = \arg\min_{A,b}\ \text{CrsEnt}(Av_j + b, y_j) + \frac{\lambda_1}{C}|b|_2^2 + \frac{\lambda_2}{C^2}|\bar{D}(A)|_2^2, \quad x_j \in \mathcal{X}_{\text{val}}. \tag{6}$$

Here, v_i denotes the model logits, $A \in \mathbb{R}^{C\times C}$ is the learned scaling matrix, and $b \in \mathbb{R}^C$ is a bias vector. Compared to simple temperature scaling (*Pseudo-T*), this method introduces significantly more parameters, increasing the risk of overfitting to the validation set $\mathcal{X}_{\text{val}}$. To mitigate this, we apply regularization terms: one to penalize large bias magnitudes, and another to suppress off-diagonal entries in A (denoted $\bar{D}(A)$). In practice, this results in a flexible yet stable calibration scheme: the diagonal entries of A remain expressive, while off-diagonal interactions are dampened. We refer to this pseudo-label method as *Pseudo-D*.

4 Experiments

4.1 Dataset

We use an anonymized private dataset obtained from a tertiary hospital with ethics approval. The dataset contains 2572 retrospective echo studies, acquired with Philips iE33, Vivid i, and Vivid E9 transducers. The studies were first labeled as normal/mild/moderate/severe based on spectral Doppler measurements using AS diagnostic guidelines from [3]. Consistent with prior methods [15,32,30], the moderate and severe classes are combined into a single "significant class". An experienced cardiologist selected parasternal long-axis and short-axis views from each study, resulting in a total of 9117 videos. We created training, validation and test using a randomized 80/10/10 split, ensuring no patients overlap across subsets. Each video was preprocessed by extracting approximately one heart cycle, removing background UI elements, and resizing to $16 \times 224 \times 224$.

4.2 Evaluation Procedure and Metrics

We compare models trained on the following pseudo-labels: *RT4U* [9] (Eqn. 1), *Pseudo-T*, and *Pseudo-D* on selective classification and multi-view fusion. We compare the pseudo-label approaches with *Vanilla*, a baseline cross-entropy approach; *Abstention* [6], which trains an extra network branch specializing in rejecting uncertain predictions; and *TMC* [12], which specializes in combining probabilities from multiple sources.

In selective classification, models are provided with an option to "reject" the prediction $f(x)$ based on a selection function $g(x)$ and threshold τ. The effectiveness of selective classification depends on both the accuracy of $f(x)$ and the sensitivity of $g(x)$ for identifying likely misclassifications. The coverage (Eqn. 7) and accuracy (Eqn. 8) are evaluated at different thresholds. Performance over multiple thresholds can be summarized by the Area Under Risk-Coverage Curve (AURC) [8] (Eqn. 9):

$$\mathrm{Cov}(f,g,\tau) = |x : g(x) > \tau|, \tag{7}$$

$$\mathrm{Acc}(f,g,\tau) = \frac{|(x,y) : f(x) = y, g(x) > \tau|}{|x : g(x) > \tau|}, \tag{8}$$

$$\mathrm{AURC}(f,g) = \frac{1}{|\mathcal{T}|}\sum_{\tau} \mathrm{Acc}(f,g,\tau) * \mathrm{Cov}(f,g,\tau). \tag{9}$$

We choose $g(x)$ to be tied to the softmax confidence of the predicted class, since it was shown to be successful across multiple specialized methods [7]. Traditionally the AURC measures loss; we co-opt the metric to assess balanced accuracy due to the nature of the classification task. We pre-compute the set of thresholds $\mathcal{T}$ at the percentiles of coverage ranging uniformly from 50% to 100%.

We use multi-view fusion to aggregate video-level predictions to study-level. In some studies, only one or two videos clearly show signatures of stenosis on the aortic valve. Thus, we adopt a "worst-case" aggregate strategy. If all videos in the study are predicted as normal, we average the softmax probability. Otherwise, we average only over the subset of videos predicted as abnormal.

In terms of metrics, we measure the accuracy (ACC), expected calibration error (ECE), and mean average error (MAE) with class 0 as normal, 1 as early,

Table 1. Evaluation over the test set of the aortic stenosis classification dataset. Each study consists of multiple ultrasound videos of the same patient. Models predict at the video-level, and predictions are fused into the study-level. We report metrics computed separately for each class, then averaged. Best and second-best results are **bolded** and underlined, respectively.

	Video-level				Study-level			
Method	MAE ↓	ACC ↑	ECE ↓	AURC ↑	MAE ↓	ACC ↑	ECE ↓	AURC ↑
Vanilla	.257	.765	.170	.804	.224	.790	.138	.852
Abstention [6]	.250	.769	.130	.820	.216	.794	.094	.881
TMC [12]	.252	.771	.213	.746	.205	.808	.163	.777
RT4U [9]	.246	.773	.105	.804	**.171**	**.836**	.073	.877
Pseudo-T (ours)	.234	.786	.097	**.838**	.192	.808	.089	.874
Pseudo-D (ours)	**.220**	**.787**	**.096**	.793	.180	.820	**.071**	**.885**

and 2 as significant. To account for class imbalance, we report balanced metrics by computing each metric per class and then averaging the results.

4.3 Implementation Details

For both pseudo-label generation and fitting, we model $f(.)$ using R(2+1)D [29] with Kinetics400 initialization, trained with ADAM, learning rate 1e-4 for 20 epochs. We augment the input via random rotation of ±15° and cropping with ratio 0.7. To compute y'_i, we save the logits from $\mathcal{X}_{train}$ and $\mathcal{X}_{val}$ every epoch; we tune temperature parameters γ, A and b using SGD with learning rate 0.01 and $\lambda_1 = \lambda_2 = 1$.

4.4 Results and Discussion

Table 1 compares the pseudo-labeling and specialized methods at both the video- and study-level. Overall, *Pseudo-D* performs best across most evaluation metrics. *Abstention* is effective for selective classification based on AURC, but underperforms on other metrics. *TMC* has similar accuracy and MAE to other methods, but its predictions are less well-calibrated, as the method was designed primarily for aggregation rather than uncertainty estimation.

The pseudo-label based methods may be performing better on noisy imaging modalities such as ultrasound because they reduce overfitting to difficult training examples. The difference between these methods is how the output probabilities are scaled. *Pseudo-T* improves over the baseline, but its lack of class-wise calibration is limiting. This is due to the heterogeneity in class confusion. In this instance, it is easier to distinguish between normal and early than between early and significant aortic stenosis disease. The learned calibration parameters γ^*, A^* and b^* show that off-diagonal terms and biases are near zero, with the most notable differences coming from the class-specific scaling factors,

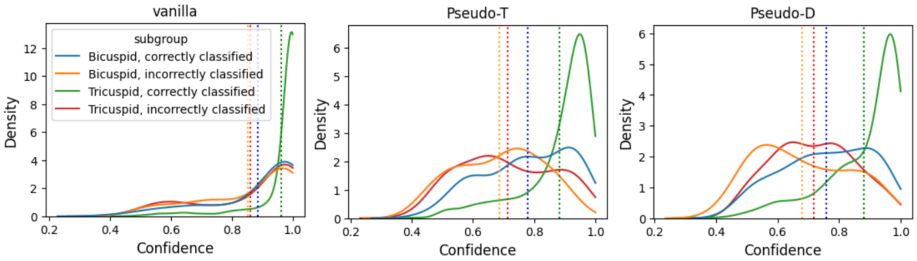

Fig. 2. Density plot of predicted confidence scores, grouped by presence of bicuspid aortic valve and prediction correctness. Dotted lines show the average confidence for each subgroup. Models trained with pseudo-label methods show stronger separation, meaning cases harder to classify correctly are more identifiable through uncertainty.

$$\gamma^* = [0.698], \quad A^* = \begin{bmatrix} 0.944 & 0.070 & -0.064 \\ -0.083 & 0.621 & 0.085 \\ 0.061 & -0.056 & 0.591 \end{bmatrix}, \quad b^* = \begin{bmatrix} -0.026 \\ 0.003 \\ 0.029 \end{bmatrix}.$$

Figure 1 presents an inference example from a particularly challenging case with significant AS. The ultrasound acquisition was difficult, resulting in low-quality videos. In acquisition 1, the valve is not clearly visible. The baseline model, trained using one-hot labels, is overconfident despite the poor image quality. In contrast, the model trained with *Pseudo-D* remains appropriately uncertain about the severity. For acquisition 2, the model's confidence increases accordingly due to the improved visual clarity.

We evaluate classification for the subset of patients with bicuspid aortic valve (BAV). The aortic valve is normally tricuspid, consisting of three leaflets. BAV is a congenital defect where two leaflets are fused, consisting of 14.1% of studies in our dataset. These cases are harder to classify due to their under-representation in the dataset and atypical valve morphology. Accurate uncertainty estimation is key to ensuring that misclassifications are flagged with high uncertainty. In Figure 2, we show the distribution of model confidence scores for each subgroup. Ideally, the confidence distributions for correct and incorrect predictions should be well separated. Since models perform worse on BAV cases, we also expect lower average confidence for this subgroup. Both pseudo-labeling methods improve subgroup separation and better align uncertainty with actual error likelihood.

5 Conclusion

We introduced a method for approximating the difficulty of training samples via NNTD, and generating pseudo-labels that reflect the unique challenge each case presents. We demonstrated improved performance in selective classification and multi-view fusion, two tasks where reliable UE is essential. NNTD can reveal valuable insights about data quality and model uncertainty, even when only a single ground truth label is available.

However, NNTD is still limited by and may vary based on the network architecture. Future work may explore distillation via creating/fitting pseudo-labels with larger/smaller networks respectively, combining dynamics-based strategies with other methods for capturing AU, accounting for inter-clinician variability through multiple label sets, and improving the sensitivity with respect to specific classes or sub-groups.

Acknowledgements. This work was supported in part by the UBC Advanced Research Computing Center (ARC), the Canadian Institutes of Health Research (CIHR), and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abe, Y.: Screening for aortic stenosis using physical examination and echocardiography. Journal of Echocardiography **19**(2), 80–85 (2021)
2. Ahmadi, N., Tsang, M., Gu, A., et al.: Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series. IEEE Transactions on Medical Imaging **43**(1), 366–376 (2024)
3. Bonow, R.O., Carabello, B.A., Chatterjee, K., et al.: Acc/aha 2006 guidelines for the management of patients with valvular heart disease: a report of the american college of cardiology/american heart association task force on practice guidelines. Journal of the American College of Cardiology **48**(3), e1–e148 (2006)
4. Chen, P., Ye, J., Chen, G., et al.: Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11442–11450 (2021)
5. Dai, W., Nazzari, H., Namasivayam, M., et al.: Identifying aortic stenosis with a single parasternal long-axis video using deep learning. Journal of the American Society of Echocardiography **36**(1), 116–118 (2023)
6. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
7. Feng, L., Ahmed, M.O., Hajimirsadeghi, H., et al.: Towards better selective classification. In: International Conference on Learning Representations (2023)
8. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. Advances in Neural Information Processing Systems **30** (2017)
9. Gu, A.N., Tsang, M., Vaseli, H., et al.: Reliable multi-view learning with conformal prediction for aortic stenosis classification in echocardiography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 327–337. Springer (2024)

10. Gulič, T.G., Makuc, J., Prosen, G., Dinevski, D.: Pocket-size imaging device as a screening tool for aortic stenosis. Wiener Klinische Wochenschrift **128**, 348–353 (2016)
11. Guo, X.: Predicting Aortic Stenosis Severity using Deep Learning. Ph.D. thesis, Massachusetts Institute of Technology (2021)
12. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2551–2566 (2022)
13. Holste, G., Oikonomou, E.K., Mortazavi, B.J., et al.: Severe aortic stenosis detection by deep learning applied to echocardiography. European Heart Journal **44**(43), 4592–4604 (2023)
14. Huang, L., Zhang, C., Zhang, H.: Self-adaptive training: beyond empirical risk minimization. Advances in Neural Information Processing Systems **33**, 19365–19376 (2020)
15. Huang, Z., Long, G., Wessler, B., Hughes, M.C.: A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In: Machine Learning for Healthcare Conference. pp. 614–647. PMLR (2021)
16. Huang, Z., Yu, X., Wessler, B.S., Hughes, M.C.: Semi-supervised multimodal multi-instance learning for aortic stenosis diagnosis. arXiv preprint arXiv:2403.06024 (2024)
17. Krishna, H., Desai, K., Slostad, B., et al.: Fully automated artificial intelligence assessment of aortic stenosis by echocardiography. Journal of the American Society of Echocardiography **36**(7), 769–777 (2023), 34th ASE Annual Scientific Sessions
18. Kull, M., Perello Nieto, M., Kängsepp, M., et al.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems **32** (2019)
19. Morvant, E., Habrard, A., Ayache, S.: Majority vote of diverse classifiers for late fusion. In: Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings. pp. 153–162. Springer (2014)
20. Nemchyna, O., Soltani, S., Solowjowa, N., et al.: Validity of visual assessment of aortic valve morphology in patients with aortic stenosis using two-dimensional echocardiography. The International Journal of Cardiovascular Imaging **37**, 813–823 (2021)
21. Otto, C., Nishimura, R., Bonow, R., et al.: 2020 acc/aha guideline for the management of patients with valvular heart disease: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. Journal of the American College of Cardiology **77**(4), e25–e197 (2021)
22. Pibarot, P., Dumesnil, J.G.: Improving assessment of aortic stenosis. Journal of the American College of Cardiology **60**(3), 169–180 (2012)
23. Pleiss, G., Zhang, T., Elenberg, E.R., et al.: Identifying mislabeled data using the area under the margin ranking. In: Advances in Neural Information Processing Systems (2020)
24. Rabanser, S., Thudi, A., Hamidieh, K., Dziedzic, A., Papernot, N.: Selective classification via neural network training dynamics. arXiv preprint arXiv:2205.13532 (2022)
25. Satopää, V.A., Baron, J., Foster, D.P., et al.: Combining multiple probability predictions using a simple logit model. International Journal of Forecasting **30**(2), 344–356 (2014)

26. Seedat, N., Crabbe, J., Bica, I., et al.: Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. In: Advances in Neural Information Processing Systems (2022)
27. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems **31** (2018)
28. Thaden, J.J., Nkomo, V.T., Lee, K.J., Oh, J.K.: Doppler imaging in aortic stenosis: the importance of the nonapical imaging windows to determine severity in a contemporary cohort. Journal of the American Society of Echocardiography **28**(7), 780–785 (2015)
29. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
30. Vaseli, H., Gu, A.N., Tsang, M.Y., et al.: Protoasnet: Comprehensive evaluation and enhanced performance with uncertainty estimation for aortic stenosis classification in echocardiography. Medical Image Analysis **103**, 103600 (2025)
31. Wang, L., Ding, Z., Tao, Z., et al.: Generative multi-view human action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6212–6221 (2019)
32. Wessler, B.S., Huang, Z., Long Jr, G., et al.: Automated detection of aortic stenosis using machine learning. Journal of the American Society of Echocardiography (2023)
33. Zhang, Q., Wu, H., Zhang, C., et al.: Provable dynamic fusion for low-quality multimodal data. In: International Conference on Machine Learning. pp. 41753–41769. PMLR (2023)