Weakly supervised classification of rare aortic valve malformations using unlabeled cardiac MRI sequences

Jason A. Fries[1,4*], Paroma Varma[2], Vincent S. Chen[1], Ke Xiao[3], Heliodoro Tejeda[3], Priyanka Saha[3], Jared Dunnmon[1], Henry Chubb[3], Shiraz Maskatia[3], Madalina Fiterau[1], Scott Delp[5], Euan Ashley[6‡], Christopher Ré[1‡], James R. Priest[3‡]

1 Department of Computer Science, Stanford University, Stanford, CA, USA
2 Department of Electrical Engineering, Stanford University, Stanford, CA, USA
3 Department of Pediatrics, Stanford University, Stanford, CA, USA
4 Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
5 Department of Bioengineering, Stanford University, Stanford, CA, USA
6 Department of Medicine, Stanford University, Stanford, CA, USA

‡These authors share senior authorship.
¤Current Address: Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
* jason-fries@stanford.edu

Abstract

Recent releases of population-scale biomedical repositories such as the UK Biobank have enabled unprecedented access to prospectively collected medical imaging data. Applying machine learning methods to analyze these data holds great promise in facilitating new insights into the genetic and epidemiological associations between anatomical structures and human health. However, the majority of these imaging data are unlabeled and deriving insights is hindered by the cost of manually annotating data at sufficient scale to train state-of-the-art deep learning models. In this work, we develop a weakly supervised deep learning model for Bicuspid Aortic Valve (BAV) classification using up to 4,000 unlabeled cardiac MRI sequences, comprising a total of 120,000 images. Instead of requiring manually labeled training data, weak supervision relies on noisy heuristic functions defined by domain experts to automatically generate large-scale, imperfect training sets. By leveraging new theoretical work on coping with label noise, models can use weaker supervision sources than was previously possible. In our BAV models, this approach substantially outperforms a traditional supervised baseline trained on hand-labeled data alone, with a 64% improvement in mean F1 score (37.8 to 61.4) on held out test data. In a validation experiment using 9,230 individuals with MRIs and long-term outcome data from the UK Biobank, applying the best-performing BAV classification model identified a subset of individuals with a 1.8-fold increase in risk of a major adverse cardiac event ($p < 0.001$). This work formalizes the first deep learning baseline for aortic valve classification and outlines a general strategy for using weak supervision to analyze large collections of unlabeled medical images.

Author summary

We developed a deep learning model for Bicuspid Aortic Valve (BAV) classification using up to 4,000 unlabeled cardiac MRI sequences, comprising a total of 120,000 images. Instead of requiring manually labeled training data, as is typical in machine learning, our approach relies on noisy heuristic functions defined by domain experts to automatically generate large-scale, imperfect training sets. In our experiments, this approach substantially outperforms a traditional supervised baseline trained on hand-labeled data alone. In a validation experiment using 9,230 individuals with MRIs and long-term outcome data from the UK Biobank, applying the best-performing BAV classification model identified a subset of individuals with a 1.8-fold increase in risk of a major adverse cardiac event. This work formalizes the first deep learning baseline for aortic valve classification and outlines a general strategy for using weak supervision to analyze large collections of unlabeled medical images.

Introduction

Bicuspid Aortic Valve (BAV) is the most common congenital malformation of the heart, occurring in 0.5-2% of the general population [1] and is associated with a variety of poor health outcomes [2]. In isolation, valvular dysfunction in BAV often leads to substantial cardiovascular pathology requiring surgical replacement of the aortic valve [3]. Machine learning models for automatically identifying aortic valve malformations via medical imaging could enable new insights into genetic and epidemiological associations with cardiac morphology. However, our understanding of the etiologies of BAV and its disease correlates are limited by the variability in age of diagnosis and the absence of large, prospectively collected imaging datasets.

Recently, the UK Biobank released a dataset of >500,000 individuals with comprehensive medical record data prior to enrollment along with long-term followup. Importantly these data also include prospectively obtained medical imaging and genome-wide genotyping data on 100,000 participants [4], including the first 14,328 subject release of phase-contrast cardiac magnetic resonance imaging (MRI) sequences. Phase-contrast cardiac MRI sequences are multi-view video clips that measure blood flow. Their high-dimensionality and overall complexity makes them appealing candidates for use with deep learning [5]. However, these prospectively collected MRIs are unlabeled, and the low prevalence of BAV introduces considerable challenges in building labeled datasets at the scale required to train deep learning models.

Obtaining labeled training data is one of the largest practical roadblocks to building deep learning models for use in medicine [6]. State-of-the-art deep learning models obviate manual feature engineering [7] by learning features directly from labeled data. However, constructing massive labeled datasets is a time-consuming and expensive process. Recent deep learning efforts in medical imaging for detecting diabetic retinopathy [8] and cancerous skin lesions [9] each required 130,000 labeled images generated by up to 7 ophthalmologists and 21 dermatologists. Standard scalable labeling approaches such as crowdsourcing are often unsuitable for medical datasets due to the domain expertise required to assign labels and the logistics of working with protected health information. More fundamentally, labels are static artifacts with sunk costs: labels themselves do not transfer to different datasets and changes to annotation guidelines necessitate re-labeling data.

An alternative to manual labeling is *weak supervision*, a collection of approaches that leverage noisy or indirect labels to train machine learning models. Several recent weak supervision frameworks use a generative model to encode domain knowledge provided in the form of noisy heuristics or *labeling functions* [10, 11] which are applied to unlabeled data to generate imperfect training data. This approach leverages unlabeled data to estimate the unobserved accuracies of labeling sources as well as infer complex statistical dependencies among labeling functions [12, 13]. The resulting generative model is then applied to unlabeled data to produce probabilistic labels, which are used to train a discriminative model such as a deep neural network. The deep learning model then learns rich feature representations from the input data, allowing it to generalize beyond the heuristics encoded in labeling functions. Unlike labels, labeling functions are easily modified and shared across datasets, providing a flexible method for generating and refining labeled datasets at the scale required to train deep learning models.

Weakly supervised machine learning methods are promising for cardiac medical imaging, a speciality that poses many computational challenges. The heart is a dynamic anatomical structure with chambers and valves, each moving in 3 dimensions with periodicity that may range from 1 to 3 Hz depending on age and health status. Cardiac imaging entails complex manual alignment to cardiac structures and the capture of multiple sequences coordinated to cardiac cycle and patient respiration. Due to the complexity of imaging output and need for human interpretation, studies utilizing cardiac MRI are mostly limited to single centers relying on human readers of clinically obtained data for functional information. For these reasons, obtaining large-scale labeled data for the space of possible cardiac pathologies is especially challenging.

We build on recent weak supervision techniques to train a state-of-the-art hybrid Convolutional Neural Network / Long Short Term Memory (CNN-LSTM) model for BAV classification. Our pipeline closely matches a realistic application setting, where we combine a small set of hand-labeled data with a large repository of unlabeled MRI sequences from the UK Biobank. This approach allows us to train deep learning models without manually constructing massive labeled datasets, substantially lowering the time and cost required to construct state-of-the-art imaging models.

Finally, to assess the real-world relevance of our image classification model, we applied the CNN-LSTM to a cohort of 9,230 new patients with long-term outcome data and MRIs from the UK Biobank. For patients identified as having BAV we found a 1.8-fold increase in risk of a major adverse cardiac event. Our approach demonstrates how real-world health outcomes can be learned directly from large-scale, unlabeled medical imaging data.

Materials and methods

Dataset

From 2006-2010, the UK Biobank recruited 502,638 participants aged 37-73 years in an effort to create a comprehensive, publicly available health-targeted dataset. The initial release of UK Biobank imaging data includes cardiac MRI sequences for 14,328 subjects [14], including eight cardiac imaging sets. Three sequences of phase-contrast MRI images of the aortic valve registered in an en face view at the sinotubular junction were obtained. Fig 1 shows example MRI videos in all encodings: raw anatomical images (CINE); magnitude (MAG); and velocity encoded (VENC) [15]. Video examples are available in S1 Videos. In MAG and VENC series, pixel intensity directly maps to velocity of blood flow. This is performed by exploiting the difference in phase of the transverse magnetism of protons within blood when flowing parallel to a gradient magnetic field, where the phase difference is proportional to velocity. CINE images encode anatomical information without capturing blood flow. All phase contrast MRI sequences are 30 frames, 12-bit grayscale color, and 192 x 192 pixels.

Fig 1. Example MRI sequence data for BAV and TAV subjects. (*Top*) Uncropped MRI frames for CINE, MAG, and VENC series in an oblique coronal view of the thorax centered upon an en face view of the aortic valve at sinotubular junction (red boxes). (*Middle*) 15-frame subsequence of a phase-contrast MRI for all series, with peak frame outlined in blue. (*Bottom*) MAG frames at peak flow for 12 patients, broken down by class: (*left*) bicuspid aortic valve (BAV) and (*right*) tricuspid aortic valve (TAV).

MRI preprocessing

All MRIs were preprocessed to: (1) localize the aortic valve to a 32x32 crop image size; and (2) align all image frames by peak blood flow in the cardiac cycle. Since the MAG series directly captures blood flow —and the aorta typically has the most blood flow—both of these steps are straightforward using standard threshold-based image processing techniques when the series is localized to a cross-sectional plane at the sinotubular junction. Selecting the pixel region with maximum standard deviation across all frames localized the aorta, and selecting the frame with maximum z-score identified peak blood flow. See S1 Appendix for implementation details. Both heuristics were very accurate (>95% as evaluated on the development set) and selecting a ±7 frame window around the peak frame f_{peak} captured 99.5% of all aorta variation. All three MRI series were aligned to this peak before classification.

Gold standard annotations

We created a set of gold standard labels for 412 patients (12,360 individual MRI frames): a *development* set (100 controls and 6 BAV patients); a *validation* set (208 controls and 8 BAV patients); and a held-out *test* set (88 controls and 3 BAV patients). The development set was selected via chart review of administrative codes (ICD9, ICD10, or OPCS4) consistent with BAV and followed by manual annotation. The validation and test sets were sampled at random with uniform probability from all 14,328 MRI sequences to capture the distribution of classes expected at test time. Development and validation set MRIs were annotated by a single cardiologist (JRP). All test set MRIs were annotated by 3 cardiologists (JRP, HC, SM) and final labels were assigned based on a majority vote across annotators. For inter-annotator agreement on the test set, Fleiss's Kappa statistic was 0.354. This reflects a fair level of agreement amongst annotators given the difficulty of the task. Test data was withheld during all aspects of model development and used solely for the final model evaluation.

Weak supervision

There is considerable research on using indirect or weak supervision to train machine learning models without relying entirely on manually labeled data [10,16,17]. One longstanding approach is *distant supervision* [18,19], where indirect sources of labels are used to to generate noisy training instances from unlabeled data. For example, in the ChestX-ray8 dataset [20] disorder labels were extracted from clinical assessments found in radiology reports. Unfortunately, we often lack access to patient notes or, as in the case of BAV, the class of interest itself may be rare and underdiagnosed in existing medical records. Another strategy is to generate noisy labels via crowdsourcing [21,22], which in some medical imaging tasks can perform as well as trained experts [23,24]. In practice, however, crowdsourcing is logistically difficult when working with protected health information such as MRIs. A significant challenge in all weakly supervised approaches is correcting for label noise, which can negatively impact end model performance. Noise is commonly addressed using rule-based and generative modeling strategies for estimating the accuracy of label sources [25,26].

In this work, we use the recently proposed *data programming* [10] method, a generalization of distant supervision that uses a generative model to learn both the unobserved accuracies of labeling sources and statistical dependencies between those sources [12,13]. In this approach, source accuracy and dependencies are estimated without requiring labeled data, enabling the use of weaker forms of supervision to generate training data, such as using noisy heuristics from clinical experts. For example, in BAV patients the phase-contrast imaging of flow through the aortic valve has a distinct ellipse or asymmetrical triangle appearance compared to the more circular aorta in TAV patients. This is the reasoning a human might apply when directly examining an MRI. In data programming these types of broad, often imperfect domain insights are encoded into functions that vote on the potential class label of unlabeled data points. This allows us to weakly supervise tasks where indirect label sources, such as patient notes with assessments of BAV, are not available.

The idea of encoding domain insights is formalized as *labeling functions* —black box functions which vote on unlabeled data points. The only restriction on labeling functions is that they vote correctly with probability better than random chance. The output of these labeling functions is used to learn a generative model of the underlying annotation process, where each labeling function is weighted by its estimated accuracy to generate probabilistic, training labels $y_i \in [0, 1]$. These probabilistically labeled data are then used to train an off-the-shelf discriminative model such as a deep convolutional neural network. Critically, the final discriminative model learns features from the entire MRI sequence, rather than the restricted space of inputs used by labeling functions. This allows the model to generalize beyond the heuristics encoded in labeling functions.

Generative model

In our setting, patient MRIs are represented as a collection of m frames $X = \{x_1, ..., x_m\}$. Each frame is modeled as an unlabeled data point x_i and latent random variable $y_i \in \{-1, 1\}$, corresponding to the true (unobserved) frame label. Supervision is provided as a set of n labeling functions $\lambda_1, ..., \lambda_n$ that define a mapping $\lambda_j : x_i \to \Lambda_{ij}$ where $\Lambda_{i1}, ..., \Lambda_{in}$ is the vector of labeling function votes. In binary classification, Λ_{ij} is in the domain $\{-1, 0, 1\}$, i.e., *false*, *abstain*, and *true*, resulting in a label matrix $\Lambda \in \{-1, 0, 1\}^{m \times n}$.

The relationship between unobserved labels y and the label matrix Λ is modeled using a factor graph [27]. We learn a probabilistic model that best explains $\hat{\Lambda}$, the empirical matrix observed by applying labeling functions to unlabeled data. In the basic data programing model, this consists of n accuracy factors between $\lambda_1, ..., \lambda_n$ and y

$$\phi_j^{Acc}(\Lambda_i, y_i) := y_i \Lambda_{ij} \tag{1}$$

Other dependencies among labeling functions (e.g., pairwise similarities) can be learned by defining additional factors. These factors may be specified manually or inferred directly from unlabeled data. The generative model's factor weights θ are estimated by minimizing the negative log likelihood of $p_\theta(\hat{\Lambda})$ using contrastive divergence [28]. Optimization is done using standard stochastic gradient descent with Gibbs sampling for gradient estimation.

Learning dependencies automatically from unlabeled data is critical in imaging tasks, where labeling functions are dependent on a complex space of domain features or *primitives*, including edges, textures, and semantic objects such as segmentations of anatomical structures. We use the generative model enhancements proposed in Varma et al. [12] to infer higher order dependency structure between labeling functions based on their interactions with domain primitives. This approach requires defining a space of feature primitives (e.g., the area of a segmentation mask) that serves as an additional input to the generative model. These features can come from any source, but in this work we use simple shape statistics and pixel intensity values.

The final weak supervision pipeline requires two inputs: (1) primitive feature matrix; and (2) observed label matrix $\hat{\Lambda}$. For generating $\hat{\Lambda}$, we take each patient's frame sequence $\bar{x}_i = \{x_{1i}, ...x_{30i}\}$ and apply labeling functions to a window of t frames $\{x_{(f_{peak}-t/2)}, ..., x_{(f_{peak}+t/2)}\}$ centered on f_{peak} , i.e., the frame mapping to peak blood flow. Here $t = 6$ performed best in our generative model experiments. The output of the generative model is a set of *per frame* probabilistic labels $\{y_1, ..., y_m\}$ where $m = t \times N$, the number of patients . To compute a single, *per patient* probabilistic label, $\bar{y}_i$, we assign the mean probability of all t patient frames if $mean(\{y_{1i}, ..., y_{ti}\}) > 0.9$ and the minimum probability if $min(\{y_{1i}, ..., y_{ti}\}) < 0.5$. Patient MRIs that did not meet these thresholds, 7% (304/4543), were removed from the final weak label set. The final weakly labeled training consists of all MRI frames and per-patient labels: $\hat{X} = \{\bar{x}_i, ..., \bar{x}_N\}$ and $\hat{Y} = \{\bar{y}_i, ..., \bar{y}_N\}$.

Extracting domain primitives

All primitives are generated using a binarized segmentation mask of the aortic valve for each frame in a patient's MAG series. Since the generative model can handle noisy labeling functions and primitives, we use simple threshold techniques such as Otsu's method [29] to generate binary segmentation masks. All masks were used to compute primitives for: (1) *area*; (2) *perimeter*; (3) *eccentricity* (a [0,1) measure comparing the mask shape to an ellipse, where 0 indicates a perfect circle); (4) *pixel intensity* (the mean pixel value for the entire mask); and (5) *ratio* (the ratio of area over perimeter squared). Since the size of the heart and anatomical structures correlate strongly with patient sex, we normalized these features by two population means stratified by sex in the unlabeled set. All image preprocessing was computed using scikit-image [30].

Designing labeling functions

Primitives define semantic abstractions over raw pixel data and allow domain experts to more easily encode heuristics for BAV classification using labeling functions. For example, geometric features capture the domain intuition that the prototypical TAV case has a symmetrical circular or triangular shape while BAV has the appearance of an ellipse or asymmetrical triangle. The ratio feature, on the other hand, is a size-independent measure of circularity, and BAV valves tend to be non-circular. These insights are cumbersome to encode using pixel information alone.

We designed 5 labeling functions using the primitives described above. For model simplicity, labeling functions were restricted to using threshold-based, frame-level information for voting. All labeling function thresholds were selected manually using distributional statistics computed over all primitives for the expert-labeled development. (See S1 Table for complete labeling function implementations). The complete weak supervision pipeline is shown in Fig 2.

Fig 2. Weak supervision workflow. Pipeline for probabilistic training label generation based on user-defined primitives and labeling functions. Primitives and labeling functions (step 1) are used to weakly supervise the BAV classification task and programmatically generate probabilistic training data from large collections of unlabeled MRI sequences (step 2), which are then used to train a noise-aware deep learning classification model (step 3).

Discriminative model

Our discriminative model classifies BAV/TAV status using t contiguous MRI frames ($5 \leq t \leq 30$, where t is a hyperparameter) and a single probabilistic label per patient. This model consists of two components: a *frame encoder* for learning frame-level features and a *sequence encoder* for combining individual frames into a single feature representation vector. For the frame encoder, we use a Dense Convolutional Network (DenseNet) [31] with 40 layers and a growth rate of 12, pretrained on 50,000 images from CIFAR-10 [32]. We tested two other common pretrained image neural networks (VGG16 [33], ResNet-50 [34]), but found that a DenseNet40-12 model performed best overall, aligning with previous reports [31]. The DenseNet architecture takes advantage of low-level feature maps at all layers, making it well-suited for medical imaging applications where low-level features (e.g., edge detectors) often carry substantial explanatory power.

For the sequence encoder, we used a Bidirectional Long Short-term Memory (LSTM) [35] sequence model with soft attention [36] to combine all MRI frame representations. Soft attention provides a fully differentiable layer for optimizing the weighted mean of frame representations, allowing the network to automatically identify the most informative frames in MRI sequences. We explored simpler feature pooling architectures (e.g, mean/max pooling), but each of these methods was outperformed by the LSTM in our experiments. The final hybrid CNN-LSTM architecture aligns with recent methods for state-of-the-art video classification [37,38] and 3D medical imaging [39].

The CNN-LSTM model is trained using noise-aware binary cross entropy loss L:

$$\hat{w} = argmin_w \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{y \sim \hat{Y}} [L(w, x_i, y)] \tag{2}$$

This is analogous to standard supervised learning loss, except we are now minimizing the expected value with respect to $\hat{Y}$ [10]. This loss enables the discriminative model to take advantage the more informative probabilistic labels produced by the generative model. Fig 3 shows the complete discriminative model pipeline.

Fig 3. Deep neural network for MRI sequence classification. Each MRI frame is encoded by the DenseNet into a feature representation f_{xi}. These frame features are fed in sequentially to the LSTM sequence encoder, which uses a soft attention layer to learn a weighted mean of all frames S_{emb}. This forms the final representation used for binary classification

Training and hyperparameter tuning

The development set was used to write all labeling functions and the validation set was used for all model hyperparameter tuning. All models were evaluated with and without data augmentation. Data augmentation is typically used in deep learning models as a way to increase available training data and encode known invariances into the final model, e.g., BAV/TAV status does not change under translation so generating shifted MRI training images should not change the class label. We used a combination of crops and affine transformations commonly used by state-of-the-art image classifiers [40]. We tested models using all 3 MRI series (CINE, MAG, VENC with a single series per channel) and models using only the MAG series. The MAG series performed best, so we only report those results here.

Hyperparameters were tuned for L2 penalty, dropout, learning rate, and the representation size of our last hidden layer before classification. Augmentation hyperparameters were tuned to determine final translation, rotation, and scaling ranges. All models use validation-based early stopping with F1 score as the stopping criterion. The probability threshold for classification was tuned using the validation set for each run to address known calibration issues when using deep learning models [41]. Architectures were tuned using a random grid search over 10 models for non-augmented data and 24 for augmented data.

Evaluation metrics

Classification models were evaluated using *positive predictive value* (precision), *sensitivity* (recall), *F1 score* (i.e., the harmonic mean of precision and recall), and *area under the ROC curve* (AUROC). Due to the extreme class imbalance of this task we also report *discounted cumulative gain* (DCG) to capture the overall ranking quality of model predictions [42]. Each BAV or TAV case was assigned a relevance weight r of 1 or 0, respectively, and test set patients were ranked by their predicted probabilities. DCG is computed as $\sum_i^n \frac{r_i}{log_r(i+1)}$ where n is the total number of instances and i is the corresponding rank. This score is normalized by the DCG score of a perfect ranking (i.e., all true BAV cases in the top ranked results) to compute normalized DCG (NDCG) in the range [0.0,1.0]. Higher NDCG scores indicate that the model does a better job of ranking BAV cases higher than TAV cases.

All scores were computed using test set data, using the best performing models found during grid search, and reported as the mean and 95% confidence intervals of 5 different random model weight initializations.

For labeling functions, we report two additional metrics: *coverage* (Eq. 3) a measure of how many data points a labeling function votes $\{-1, 1\}$ on; and *conflict* (Eq. 4) the percentage of data points where a labeling function disagrees with one or more other labeling functions.

$$coverage_{\lambda_j} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\lambda_j(x_i) \in \{-1, 1\}) \qquad (3)$$

$$conflict_{\lambda_j} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\sum_{k\neq j}^{\lambda_n}\mathbb{1}(\lambda_j(x_i) \in \{-1,1\} \wedge \lambda_j(x_i) \neq \lambda_k(x_i))) > 0 \qquad (4)$$

Model evaluation with clinical outcomes data

To construct a real-world validation strategy dependent upon the accuracy of image classification but completely independent of the imaging data input, we used model-derived classifications (TAV vs. BAV) as a predictor of validated cardiovascular outcomes using standard epidemiological methods. For 9,230 patients with prospectively obtained MRI imaging who were excluded from any aspect of model construction, validation, or testing we performed an ensemble classification with the best performing CNN-LSTM model.

For evaluation we assembled a standard composite outcome of *major adverse cardiovascular events* (MACE). Phenotypes for MACE were inclusive of the first occurrence of coronary artery disease (myocardial infarction, percutaneous coronary intervention, coronary artery bypass grafting), ischemic stroke (inclusive of transient ischemic attack), heart failure, or atrial fibrillation. These were defined using ICD-9, ICD-10, and OPCS-4 codes from available hospital encounter, death registry, and self-reported survey data of all 500,000 participants of the UK Biobank at enrollment similar to previously reported methods [43].

Starting 10 years prior to enrollment in the study, median follow up time for the participants with MRI data included in the analysis was 19 years with a maximum of 22 years. For survival analysis, we employed the "survival" and "survminer" packages in R version 3.4.3, using aortic valve classification as the predictor and time-to-MACE as the outcome, with model evaluation by a simple log-rank test.

To verify the accuracy of the CNN-LSTM's predicted labels, 36 MRI sequences (18 TAV and BAV patients) were selected randomly for review by a single annotator (JRP). The output of the last hidden layer was visualized using a t-distributed stochastic neighbor embedding (t-SNE) [44] plot to assist error analysis.

Results

Baseline models

We compare our weakly supervised models against two traditionally supervised baselines using identical CNN-LSTM architectures: (1) expert labels alone and (2) expert labels with data augmentation. In these experiments, the expert labeled development set was used as the training set. Due to class imbalance (6:100), training data was rebalanced by oversampling BAV cases with replacement.

Weak supervision performance at scale

We evaluate the impact of training set size on weak supervision performance. These models are trained using only weakly labeled training data, i.e., no hand-labeled MRIs. All probabilistic labels are split into positive and negative bins using a threshold of 0.5 and sampled uniformly at random with replacement to create balanced, training sets, e.g., sample 50 BAV and 50 TAV for a training set size of 100. We used balanced samples sizes of $\{50, 250, 500, 1000, 2000, 4000\}$. The final class balance for all 4,239 weak labels in the training set was 264/3975 BAV/TAV. Full scale-up metrics for weak labels are shown in Fig 4.

Fig 4. Weak supervision scale up performance metrics. Metrics include *positive predictive value* (precision), *sensitivity* (recall), *area under the ROC curve* (AUROC), and *normalized discounted cumulative gain* (NDCG). The y-axis is the score in [0,100] and the x-axis is the number of unlabeled MRIs used for training. The dashed horizontal line indicates the expert-labeled baseline model. Shaded regions indicate 95% confidence intervals. Mean precision increased 128% (30.7 to 70.0) using 4,239 weakly labeled MRIs; sensitivity (recall) matched performance of the expert-labeled baseline (53.3 vs. 60.0). At $\geq$ 1264 weak training examples, all models exceeded the performance of a model trained on 106 expert-labeled MRIs.

Models trained with 4,239 weak labels and augmentation performed best overall, matching or exceeding all metrics compared to the best performing baseline model, expert labels with augmentation. The best weak supervision model had a 64% improvement in mean F1 score (61.4 vs. 37.8) and 128% higher mean precision (30.7 to 70.0). This model had higher mean area under the ROC curve (AUROC) (+13%) and normalized discounted cumulative gain (NDCG) (+57%) scores. In Table 1, we report baseline model performance and the best weak supervision models found across all scale-up experiments. See S3 Fig for ROC plots across all scale-up sizes.

Table 1. Best Performing Weak Supervision Models vs. Baselines

Model	**Train Size**	**Precision [95% CI]**	**Recall [95% CI]**	**F1 [95% CI]**	**AUROC [95% CI]**	**NDCG [95% CI]**
BASELINE: Labels	106	19.5 [12.5, 28.6]	40.0 [33.3, 66.7]	26.1 [18.2, 40.0]	87.4 [70.0, 92.7]	44.4 [37.2, 50.8]
BASELINE: Labels + Augmentation	106	30.7 [20.8, 40.6]	53.3 [38.7, 68.0]	37.8 [27.7, 47.9]	83.4 [79.5, 87.3]	55.7 [51.5, 59.9]
Weak Supervision	4239	**83.3** [64.5, 100.0]	53.3 [38.7, 68.0]	60.8 [50.6, 71.0]	91.4 [87.8, 95.0]	84.5 [81.1, 88.0]
Weak Supervision + Augmentation	4239	70.0 [55.4, 84.6]	**60.0** [48.1, 72.0]	**61.4** [55.3, 67.5]	**94.4** [91.3, 97.6]	**87.3** [83.6, 91.0]

Labeling Function Scores

Table 2 shows individual labeling function performance on test data, where metrics were computed per-frame. Precision, recall, and F1 scores were calculated by counting abstain votes as TAV labels, reflecting a strong prior on TAV cases. Individually, each function was a very weak classifier with poor precision (0 - 25.0) and recall (0 - 69.1), as well as mixed coverage (9.8% - 90%) and substantial conflict with other labeling functions (8 - 41.7%). Note that labeling functions provide both negative and positive class supervision, and sometimes performed best with a specific class, e.g., LF_Intensity targets negative cases while LF_Perimeter targets positive.

Table 2. Frame-level Labeling Function Performance Metrics

Labelers	Coverage%	Conflict%	Pos. Acc.	Neg. Acc.	Precision	Recall	F1
LF_Area	22.6	11.5	76.5	62.9	25.0	31.0	27.7
LF_Perimeter	9.8	8.0	100.0	0.0	20.8	26.2	23.2
LF_Eccentricity	87.4	38.9	85.7	42.3	12.7	85.7	22.1
LF_Intensity	28.9	24.1	0.0	69.0	0.0	0.0	0.0
LF_Ratio	90.4	41.7	67.5	49.6	10.7	64.3	18.3

Orthogonal model validation using clinical outcomes data

In a time-to-event analysis encompassing up to 22 years of follow-up on the 9,230 included participants with cardiac MRI data, the individuals with model-classified BAV showed a significantly lower MACE-free survival (Hazard Ratio 1.8; 95% confidence interval 1.3-2.4, p = 8.83e-05 log-rank test) (see Fig. 5) consistent with prior knowledge of co-incidence of BAV with comorbid cardiovascular disease [45, 46]. In a linear model adjusted for age, sex, smoking, hyperlipidemia, diabetes, and BMI, individuals with model-classified BAV displayed a 2.5 mmHg increase in systolic blood pressure ($p < 0.001$).

Fig 5. Unadjusted Survival from MACE in 9,230 Participants Stratified by Model Classification. MACE occurred in 59 of 570 individuals (10.4%) classified as BAV compared to 511 of 8660 individuals (5.9%) classified as TAV over the course of a median 19 years of follow up (Hazard Ratio 1.8; 95% confidence interval 1.3-2.4, p = 8.83e-05 log-rank test).

Fig. 6 shows a t-SNE plot of BAV/TAV clusters using the CNN-LSTM's last hidden layer output (i.e., the learned representation). In the post-hoc analysis of 36 predicted MRI labels, TAV cases had 94% (17/18) PPV (precision) and BAV cases had 61% (11/18) PPV, with BAV misclassifications occurring most often in cases with visible regurgitation and turbulent blood flow.

Fig 6. Patient clustering visualization. (*Left*) t-SNE visualization of the last hidden layer outputs of the CNN-LSTM model as applied to 9,230 patient MRI sequences and (*right*) frames capturing peak flow through the aorta for a random sample of patients. Blue and orange dots represent TAV and BAV cases. The model clusters MRIs based on aortic shape and temporal dynamics captured by the LSTM. The top example box (1) contains clear TAV cases with very circular flow shapes, with (2) and (3) becoming more irregular in shape until (4) shows highly irregular flow typical of BAV. Misclassifications of BAV (red boxes) generally occur when the model fails to differentiate regurgitation of the aortic valve and turbulent blood flow through a normal appearing aortic valve orifice.

Discussion

In this work we present the first deep learning model for classifying BAV from phase-contrast MRI sequences. These results were obtained using models requiring only a small amount of labeled data, combined with a large, imperfectly labeled training set generated via weak supervision. The success of this weak supervision paradigm, especially for a classification task with substantial class-imbalance such as BAV, represents a critical first step in the larger goal of automatically labeling unstructured medical imaging from large datasets like the UK Biobank. For medical applications of machine learning as described here, we propose an additional standard of validation; that the model not only captures abnormal valve morphology, but more importantly the captured information is of real-world medical relevance. In our model, BAV individuals showed more than an 1.8-fold increase in risk for comorbid cardiovascular disease.

The current availability of large unstructured medical imaging datasets is unprecedented in the history of biomedical research, but the use of data on cardiac morphology derived from medical imaging depends upon their integration into genetic and epidemiological studies. For most aspects of cardiac structure and function, the computational tools used to perform clinical measurements require the input or supervision of an experienced user, typically a cardiologist, radiologist, or technician. Large datasets exploring cardiovascular health such as MESA and GenTAC which both include imaging data have been limited by the scarcity of expert clinical input in labeling and extracting relevant information [47, 48]. Our approach provides a scalable method to accurately and automatically label such high value datasets.

Automated classification of imaging data represents the future of imaging research. Weakly supervised deep learning tools may allow imaging datasets from different institutions which have been interpreted by different clinicians, to be uniformly ascertained, combined, and analyzed at unprecedented scale, a process termed *harmonization*. Independent of any specific research or clinical application, new machine learning tools for analyzing and harmonizing imaging data collected for different purposes will be the critical link that enables large-scale studies to connect anatomical and phenotypic data to genomic information, and health-related outcomes. For the purposes of research, such as genome-wide association studies, higher precision (positive predictive value) is more important for identifying cases. Conversely, in a clinical application the flagging of all possible cases of BAV for manual review by a clinician would be facilitated by a more sensitive threshold. The model presented here can be tuned to target either application setting.

Our analytical framework and models have limitations. Estimation of the true prevalence of uncommon conditions such as BAV and ascertainment of outcomes within a given population is complicated by classical biases in population health science. Registries of BAV typically enroll patients only with clinically apparent manifestations or treatment for disease which may not account for patients who do not come to medical attention. Estimates derived from population-based surveillance are usually limited to relatively small numbers of participants due to the cost and difficulty of prospective imaging, and small cohort sizes impede accurate estimates for rare-conditions such as BAV. Age and predisposition to research participation may also affect estimates of disease prevalence, a documented phenomenon within the UK Biobank [49]. Mortality from BAV is accrued cumulatively over time, thus studies of older participants are missing individuals with severe disease who may have died or been unable to participate.

Conversely calcific aortic valve disease, which increases in incidence with age, may result in an acquired form of aortic stenosis difficult to distinguish from BAV by cardiac flow imaging [50]. Given that the 6.2% of individuals receiving a model-classification of BAV is higher than previous population estimates of BAV prevalence (0.5 to 2%), some proportion of BAV-classified individuals almost certainly have age-related calcific aortic valve disease. Additional scrutiny of model-classified BAV cases show that the model fails to differentiate regurgitation of the aortic valve from turbulent blood flow through an aortic valve with a normal circular or symmetrically triangular appearing orifice (Fig. 6). Thus even for the current best-performing model displaying good predictive characteristics for a class-imbalanced problem, misclassification events attributable to discreet failure modes are evident for subsequent iterations of the model.

Related Work

In medical imaging, weak supervision covers a broad range of techniques using indirect or noisy labels. *Multiple instance learning* (MIL) is one common weak supervision approach in medical images [51]. Xu et al. [52] simultaneously performs binary classification and segmentation for histopathology images using a variant of MIL, where image-level labels are used to supervise both image classification and a segmentation subtask. ChestX-ray8 [20] was used in Li et al. [53] to jointly perform classification and localization using a small number of weakly labeled examples. Patient radiology reports and other medical record data are frequently used to generate noisy labels for imaging tasks [20, 54–56].

Weak supervision shares similarities with *semi-supervised learning* [57], which enables training models using a small labeled dataset combined with large, unlabeled data. The primary difference is how the structure of unlabeled data is specified in the model. In semi-supervised learning, we make smoothness assumptions and extract insights on structure directly from unlabeled data using task-agnostic properties such as distance metrics and entropy constraints [58]. Weak supervision, in contrast, relies on directly injecting domain knowledge into the model to incorporate the underlying structure of unlabeled data. In many cases, these sources of domain knowledge are readily available in existing knowledge bases, indirectly-labeled data like patient notes, or weak classification models and heuristics.

Conclusion

This work demonstrates how weak supervision can be used to train a state-of-the-art deep learning model for BAV classification using unlabeled MRI sequences. Using domain heuristics encoded as functions to programmatically generate large-scale, imperfect training data provided substantial improvements in classification performance over models trained on hand-labeled data alone. Transforming domain insights into labeling functions instead of static labels mitigates some of the challenges inherent in the domain of medical imaging, such as extreme class imbalance, limited training data, and scarcity of expert input. Most importantly, our BAV classifier could successfully identify individuals at long-term risk for cardiovascular disease, demonstrating real-world relevance of imaging models built using weak supervision techniques.

Supporting information

S1 Videos. Example MRI videos. BAV and TAV subject videos in CINE, MAG, and VENC encodings.

S1 Appendix. Aorta localization and cardiac cycle alignment. Detailed overview of MRI preprocessing.

S1 Fig. Localizing the aortic valve. (*Left*) Full, uncropped MAG series MRI frame, showing per pixel standard deviation. (*Right*) Green box is a zoom of the heart region and the red box corresponds to the aorta – the highest weighted pixel area in the image.

S2 Fig. Per-frame z-scores for a random sample of 50 MRI sequences. The majority of series only contains pixel information in the first 15 frames of data.

S3 Fig. Area under the ROC curve (AUROC) for all scaleup models. As the CNN-LSTM is trained on more weakly labeled data AUROC generally improves. In very small training set regimes (e.g., 100 - 1000 instances) using only weakly labeled data, performance degrades after > 0.6 true positive rate.

S4 Fig. Development set BAV subjects. All 6 BAV subjects used for labeling function development. For the generative model, 6 contiguous frames performed best at classifying training data using labeling functions, while in the discriminative CNN-LSTM model, 10 frames performed best. This shows how the deep learning model was better able to take advantage of subtle features at the start and end of the cardiac cycle, while labeling functions are restricted to less ambiguous features near the peak frame.

S1 Table. Complete Labeling Function Implementations.

S2 Table. CNN-LSTM Model Hyperparameter Search Grid.

Acknowledgments

We thank Seung-Pyo Lee for his initial contributions to the start of this project. We thank the Brin-Wojcicki foundation for support (EA, JRP). This work was supported in part by the Mobilize Center, a National Institutes of Health Big Data to Knowledge (BD2K) Center of Excellence supported through Grant U54EB020405 (JF, MF), the National Science Foundation (NSF) Graduate Research Fellowship under No. DGE-114747 (PV), Joseph W. and Hon Mai Goodman Stanford Graduate Fellowship (PV), and the Intelligence Community Postdoctoral Fellowship (JD). We gratefully acknowledge the support of DARPA under No. FA87501720095 (D3M), ONR under No. N000141712266 and No. N000141410102, and members of the Stanford DAWN project: Intel, Microsoft, Teradata, Google, Facebook, and VMware. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

References

1. Roberts WC, Ko JM. Frequency by decades of unicuspid, bicuspid, and tricuspid aortic valves in adults having isolated aortic valve replacement for aortic stenosis, with or without associated aortic regurgitation. Circulation. 2005;111(7):920–925.

2. Siu SC, Silversides CK. Bicuspid aortic valve disease. J Am Coll Cardiol. 2010;55(25):2789–2800.

3. Masri A, Svensson LG, Griffin BP, Desai MY. Contemporary natural history of bicuspid aortic valve disease: a systematic review. Heart. 2017;103(17):1323–1330.

4. Allen NE, Sudlow C, Peakman T, Collins R, UK Biobank. UK biobank data: come and get it. Sci Transl Med. 2014;6(224):224ed4.

5. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. npj Digital Medicine. 2018;1(1):6.

6. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 2017;21(1):4–21.

7. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012;55(10):78–87.

8. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410.

9. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118.

10. Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C. Data programming: Creating large training sets, quickly. In: Advances in Neural Information Processing Systems; 2016. p. 3567–3575.

11. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Christopher R. Snorkel: Rapid Training Data Creation with Weak Supervision. Proceedings of the VLDB Endowment. 2017;11(3):269–282.

12. Varma P, He B, Bajaj P, Banerjee I, Khandwala N, Rubin DL, et al. Inferring Generative Model Structure with Static Analysis. Adv Neural Inf Process Syst. 2017;30:239–249.

13. Bach SH, He B, Ratner A, Ré C. Learning the Structure of Generative Models without Labeled Data. 2017;70:273–282.

14. Petersen SE, Matthews PM, Bamberg F, Bluemke DA, Francis JM, Friedrich MG, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. J Cardiovasc Magn Reson. 2013;15:46.

15. Srichai MB, Lim RP, Wong S, Lee VS. Cardiovascular applications of phase-contrast MRI. AJR Am J Roentgenol. 2009;192(3):662–675.

16. Bunescu R, Mooney R. Learning to extract relations from the web using minimal supervision. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics; 2007. p. 576–583.

17. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A. Training Deep Neural Networks on Noisy Labels with Bootstrapping. Workshop contribution at ICLR. 2015; p. 1–11.

18. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol. 1999; p. 77–86.

19. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09; 2009.

20. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; p. 3462–3471.

21. Gao, Huiji and Barbier, Geoffrey and Goolsby, Rebecca. Harnessing the Crowdsourcing Power of Social Media for Disaster Relief. IEEE Intelligent Systems. 2011;26(3):10–14.

22. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision. 2017;123(1):32–73.

23. McKenna MT, Wang S, Nguyen TB, Burns JE, Petrick N, Summers RM. Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence. Med Image Anal. 2012;16(6):1280–1292.

24. Gurari D, Theriault D, Sameki M, Isenberg B, Pham TA, Purwada A, et al. How to Collect Segmentations for Biomedical Images? A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-experts, and Algorithms. In: 2015 IEEE Winter Conference on Applications of Computer Vision; 2015. p. 1169–1176.

25. Nguyen TB, Wang S, Anugu V, Rose N, McKenna M, Petrick N, et al. Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. Radiology. 2012;262(3):824–833.

26. Khetan A, Lipton ZC, Anandkumar A. Learning From Noisy Singly-labeled Data. 2017;.

27. Kschischang FR, Frey BJ, Loeliger HA. Factor graphs and the sum-product algorithm. IEEE Transactions on information theory. 2001;47(2):498–519.

28. Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Comput. 2002;14(8):1771–1800.

29. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans Syst Man Cybern. 1979;9(1):62–66.

30. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014;2:e453.

31. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;1(2):3.

32. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images; 2009.

33. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014;.

34. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

35. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.

36. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: International Conference on Machine Learning; 2015. p. 2048–2057.

37. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; p. 2625–2634.

38. Zhang K, Chao WL, Sha F, Grauman K. Video Summarization with Long Short-Term Memory. In: Computer Vision – ECCV 2016. Springer International Publishing; 2016. p. 766–782.

39. Grewal M, Srivastava MM, Kumar P, Varadarajan S. RADNET: Radiologist Level Accuracy using Deep Learning for HEMORRHAGE detection in CT Scans. IEEE Symposium on Biomedical Imaging (ISBI). 2018;.

40. Cireşan DC, Meier U, Gambardella LM, Schmidhuber J. Deep, big, simple neural nets for handwritten digit recognition. Neural Comput. 2010;22(12):3207–3220.

41. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. 2017;.

42. Järvelin K, Kekäläinen J. Cumulated Gain-based Evaluation of IR Techniques. ACM Trans Inf Syst Secur. 2002;20(4):422–446.

43. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention; 2018.

44. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 2014;.

45. Michelena HI, Desjardins VA, Avierinos JF, Russo A, Nkomo VT, Sundt TM, et al. Natural history of asymptomatic patients with normally functioning or minimally dysfunctional bicuspid aortic valve in the community. Circulation. 2008;117(21):2776–2784.

46. Koenraadt WMC, Tokmaji G, DeRuiter MC, Vliegen HW, Scholte AJHA, Siebelink HMJ, et al. Coronary anatomy as related to bicuspid aortic valve morphology. Heart. 2016;102(12):943–949.

47. Weinsaft JW, Devereux RB, Preiss LR, Feher A, Roman MJ, Basson CT, et al. Aortic Dissection in Patients With Genetically Mediated Aneurysms: Incidence and Predictors in the GenTAC Registry. J Am Coll Cardiol. 2016;67(23):2744–2754.

48. Yoneyama K, Venkatesh BA, Bluemke DA, McClelland RL, Lima JAC. Cardiovascular magnetic resonance in an adult human population: serial observations from the multi-ethnic study of atherosclerosis. J Cardiovasc Magn Reson. 2017;19(1):52.

49. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017;186(9):1026–1034.

50. Otto CM, Lind BK, Kitzman DW, Gersh BJ, Siscovick DS. Association of aortic-valve sclerosis with cardiovascular mortality and morbidity in the elderly. N Engl J Med. 1999;341(3):142–147.

51. Quellec G, Cazuguel G, Cochener B, Lamard M. Multiple-instance learning for medical image and video analysis. IEEE reviews in biomedical engineering. 2017;10:213–234.

52. Xu Y, Zhu JY, Chang EIC, Lai M, Tu Z. Weakly supervised histopathology cancer image segmentation and classification. Med Image Anal. 2014;18(3):591–604.

53. Li Z, Wang C, Han M, Xue Y, Wei W, Li LJ, et al. Thoracic Disease Identification and Localization with Limited Supervision. arXiv [csCV]. 2017;.

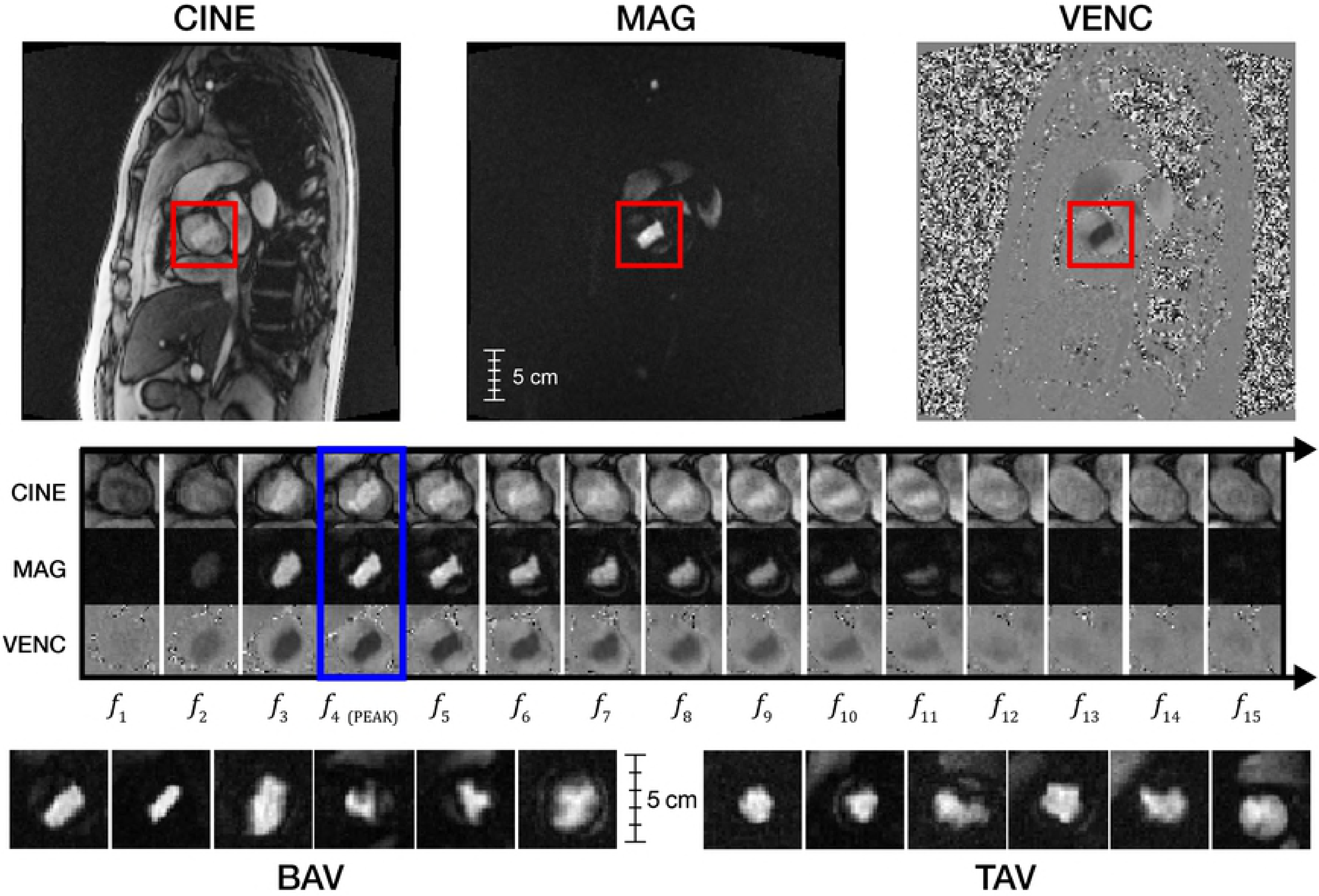
54. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. npj Digital Medicine. 2018;1(1):9.

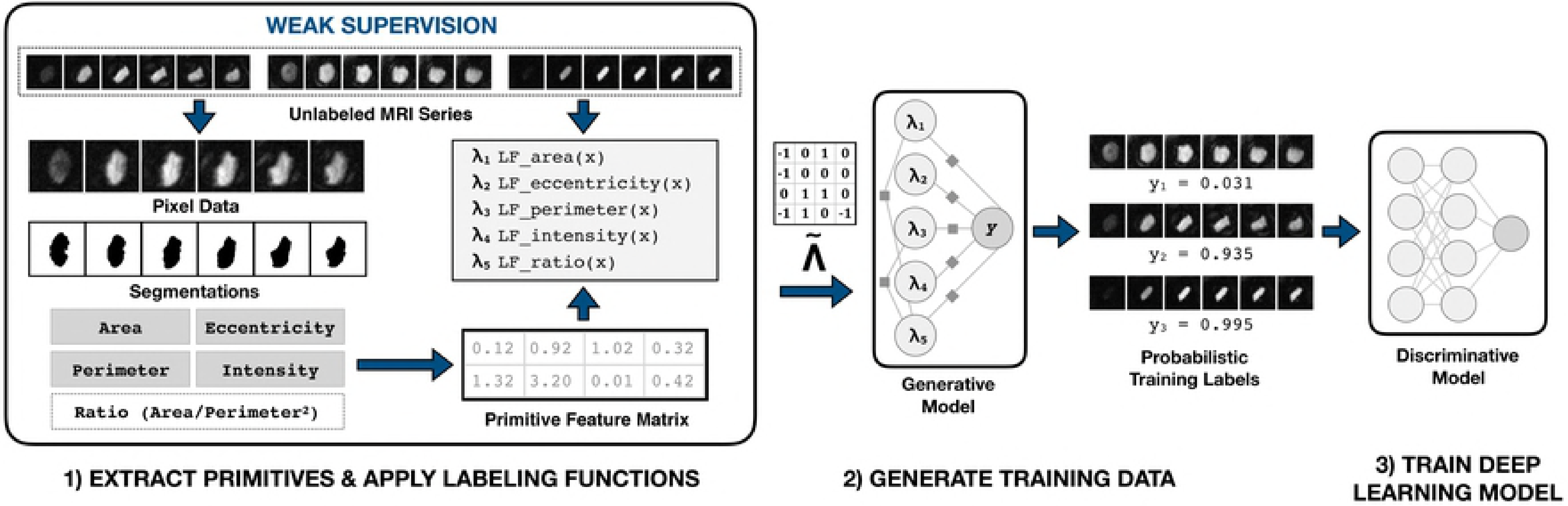
55. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv preprint arXiv:171106504. 2017;.

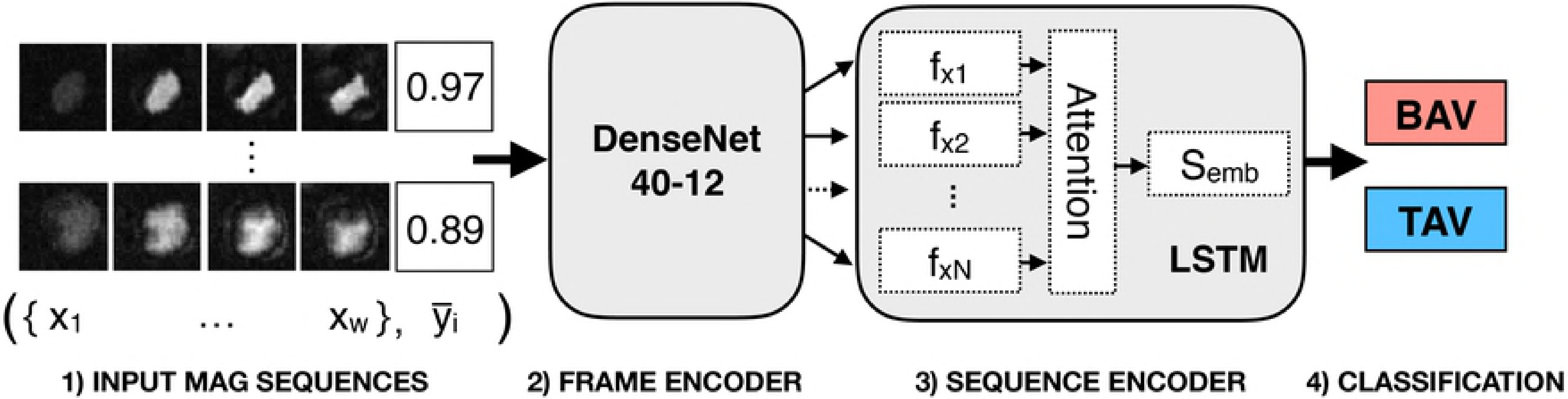
56. Wang X, Lu L, Shin HC, Kim L, Bagheri M, Nogues I, et al. Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. In: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE; 2017. p. 998–1007.

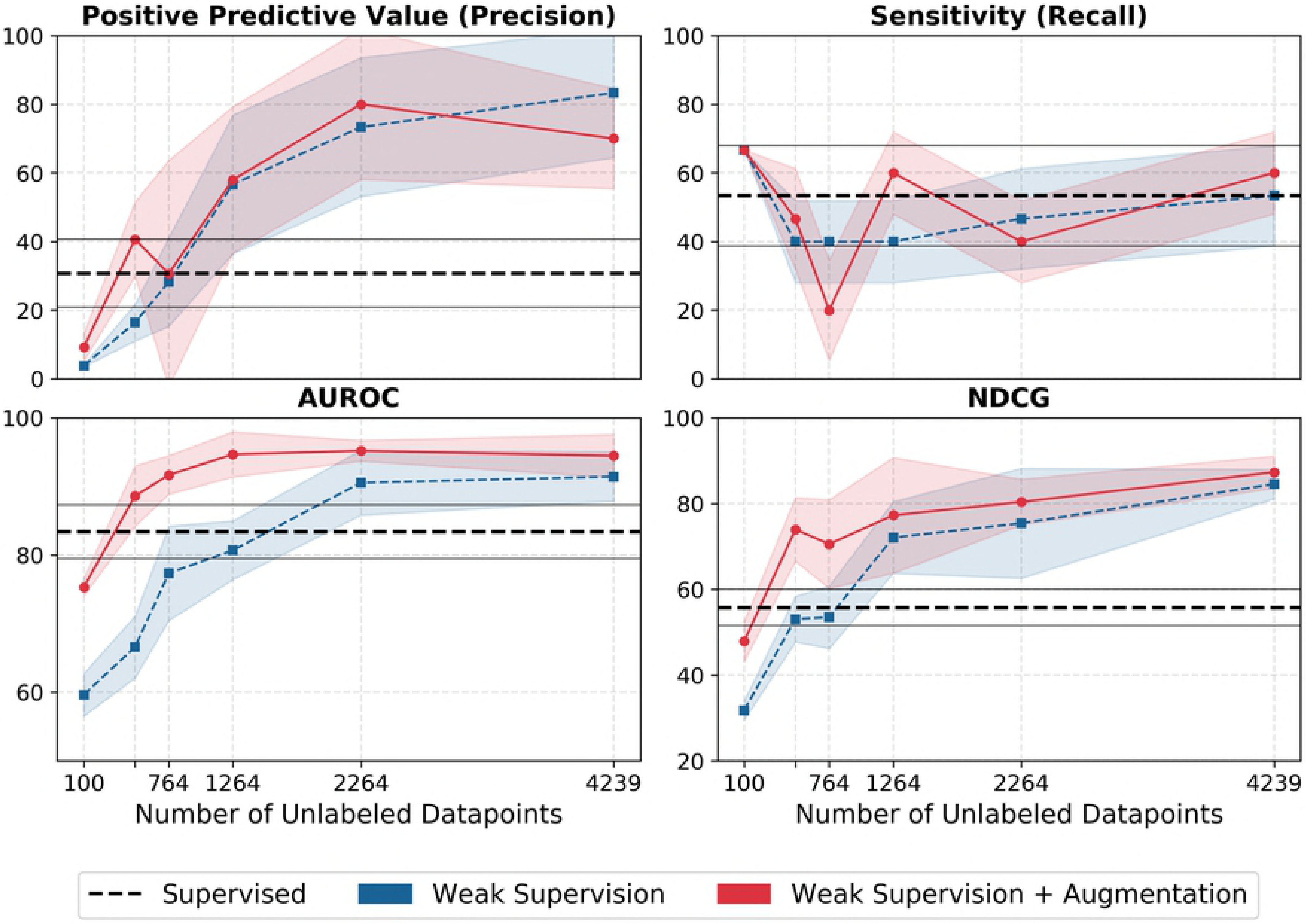
57. Semi-Supervised Learning. In: Diniz PSR, Suykens JAK, Chellappa R, Theodoridis S, editors. Academic Press Library in Signal Processing. vol. 1 of Academic Press Library in Signal Processing. Elsevier; 2014. p. 1239–1269.

58. Sun H, Cohen WW, Bing L. Semi-Supervised Learning with Declaratively Specified Entropy Constraints. 2018;.

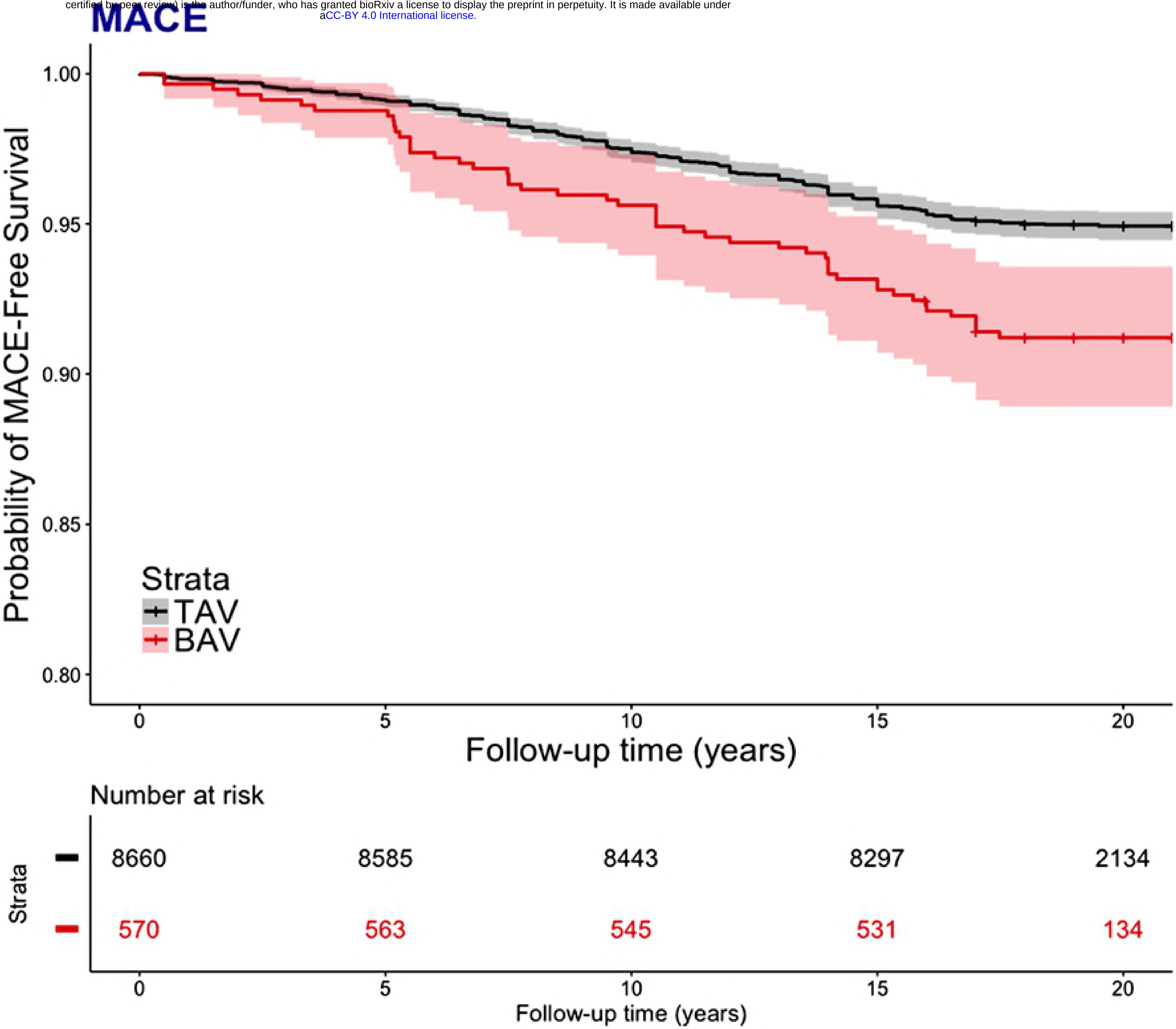

CINE
MAG
VENC
5 cm
CINE
MAG
VENC
f_1
f_2
f_3
f_4 (PEAK)
f_5
f_6
f_7
f_8
f_9
f_{10}
f_{11}
f_{12}
f_{13}
f_{14}
f_{15}
5 cm
BAV
TAV

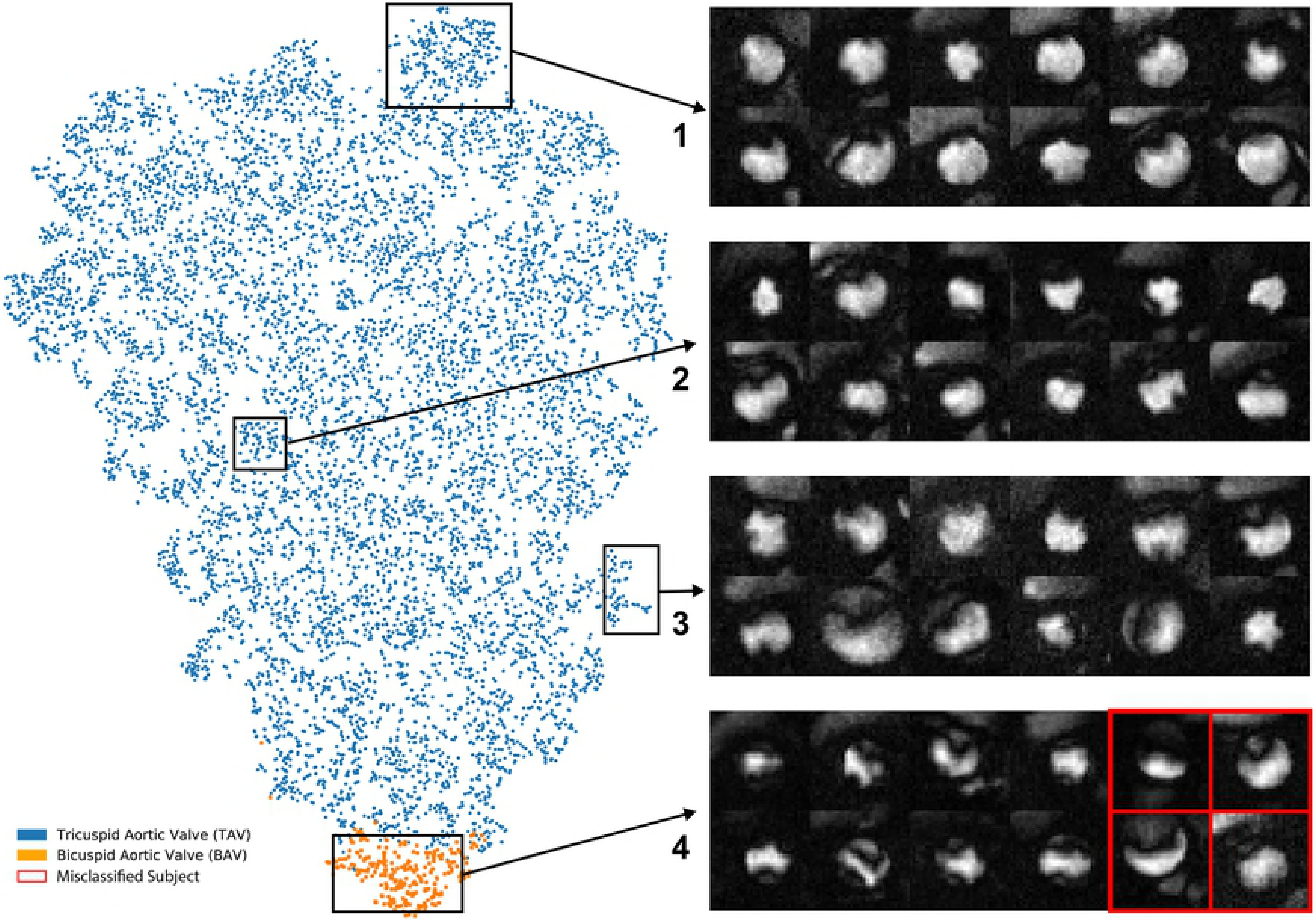

WEAK SUPERVISION
Unlabeled MRI Series
Pixel Data
Segmentations
Area
Eccentricity
Perimeter
Intensity
Ratio (Area/Perimeter²)
0.12 0.92 1.02 0.32
1.32 3.20 0.01 0.42
Primitive Feature Matrix
λ₁ LF_area(x)
λ₂ LF_eccentricity(x)
λ₃ LF_perimeter(x)
λ₄ LF_intensity(x)
λ₅ LF_ratio(x)
1) EXTRACT PRIMITIVES & APPLY LABELING FUNCTIONS
-1 0 1 0
-1 0 0 0
0 1 1 0
-1 1 0 -1
Λ̃
λ₁
λ₂
λ₃
λ₄
λ₅
y
Generative Model
y₁ = 0.031
y₂ = 0.935
y₃ = 0.995
Probabilistic Training Labels
2) GENERATE TRAINING DATA
Discriminative Model
3) TRAIN DEEP LEARNING MODEL


0.97
0.89
$(\{x_1 \; \dots \; x_w\}, \; \bar{y}_i)$
DenseNet
40-12
f_{x1}
f_{x2}
f_{xN}
Attention
S_{emb}
LSTM
BAV
TAV
1) INPUT MAG SEQUENCES
2) FRAME ENCODER
3) SEQUENCE ENCODER
4) CLASSIFICATION



Positive Predictive Value (Precision)
Sensitivity (Recall)
AUROC
NDCG
100
80
60
40
20
0
100
764
1264
2264
4239
Number of Unlabeled Datapoints
Supervised
Weak Supervision
Weak Supervision + Augmentation


MACE
Probability of MACE-Free Survival
1.00
0.95
0.90
0.85
0.80
Strata
TAV
BAV
0
5
10
15
20
Follow-up time (years)
Number at risk
Strata
8660
8585
8443
8297
2134
570
563
545
531
134
0
5
10
15
20
Follow-up time (years)


1
2
3
4
Tricuspid Aortic Valve (TAV)
Bicuspid Aortic Valve (BAV)
Misclassified Subject