A deep neural network approach to automatically diagnose aortic valve stenosis using image-metadata fusion

Team 4

Victoria Wu
83492439
vw16@student.ubc.ca

Armin Saadat
47353586
arminsdt@ece.ubc.ca

Abstract

Aortic Stenosis (AS) is a heart disease associated with the narrowing of the aortic valve opening, and is the most prevalent and deadly cardiovascular disease in developed countries. Currently, AS is diagnosed using echocardiography and Doppler ultrasound, where certain parameters given by Doppler will lead to a diagnosis class. Due to the complexity associated with this method, it is not easily accessible in many rural and developing areas, and errors in measurement from Doppler will lead to inaccurate diagnosis. Machine learning methods have been applied to AS ultrasound data to classify AS severity without using Doppler with varied levels of success. In real-world scenarios however, clinicians do not rely on the ultrasound video alone to make diagnosis. Other information such as patient's health history is used to get a better understanding of the case. The proposed research aims to combine patient metadata with ultrasound echocardiogram videos as input to neural networks for AS severity prediction, without using any Doppler measurements. To assess the effect of patient's matadata and have a fair comparison, we implement two deep neural models, one working with echo videos and one working with metadata in addition to echos. Also, in addition to the main AS severity classification task, we define a regression task for estimating the Aortic Valve Area (AVA) which is an important Doppler factor for determining AS severity. We conduct our experiments on a private dataset gathered from Vancouver General Hospital (VGH), consisting of over 9000 echos provided with patient's metadata and AS severity labels. Based on our experiments, while using metadata as input does not enhance the AS severity classification accuracy by large margin, it does not yield worse results, which shows capacity for improvement in future works. Furthermore, combining the regression task of estimating the AVA with the classification task yields the best result of 72% accuracy on the testing dataset. Our code for this project is publicly available at https://github.com/Armin-Saadat/AS_project

1 Introduction

1.1 Problem Definition

Aortic stenosis is a valvular cardiac disease that is caused by calcification of the aortic valve (AV) [2]. The calcification causes the motion of the AV to be restricted, and as a result, the valve cannot open or close properly as shown in Figure 1. This leads to a narrowing of the aorta, which restricts blood flow from the heart to the rest of the body. Treatment for AS varies based on the severity but can range from monitoring symptoms to AV replacement surgery. If left undiagnosed and untreated, severe cases of AS can lead to death; it is the most deadly valvular cardiac disease. When untreated, the 5-year mortality rate of those classified with moderate and severe AS is 56% and 67%, respectively [12]. Therefore, timely diagnoses and early intervention are crucial. Currently, AS is diagnosed using

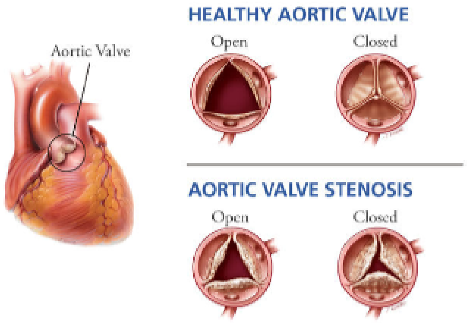

Figure 1: Aortic valve stenosis depicted as the narrowing of the aortic valve

Echocardiography (echo), an ultrasound of the heart. Doppler measurements are used to obtain 3 specific measurements: AV area (cm^2), peak velocity of the valvular jet (m/s), and mean pressure gradient ($mmHg$). These three factors are used by clinicians to determine the severity of AS, where the patient is classified as having mild, moderate, or severe AS [8]. Table 1 shows the Doppler factors' criteria for classifying AS severity.

Table 1: Classification criteria based on Doppler factors

Severity	Aortic Valve Area (cm^2)	Peak Velocity (m/s)	Mean Gradient ($mmHg$)
Mild	≥ 1.5	≤ 3	≤ 20
Moderate	$1 - 1.5$	$3 - 4$	$20 - 40$
Severe	≤ 1	≥ 4	≥ 40

Traditional ultrasound with Doppler Spectral capability can be more expensive and difficult to access for rural and lower-income communities. In addition, more training is required for the ultrasound technician to obtain Doppler measurements. To enable timely and widely available AS diagnoses, an affordable and efficient ultrasound mechanism is needed. In recent years, point-of-care ultrasound (POCUS) has become more prevalent. POCUS is a handheld ultrasound device that is capable of quickly obtaining ultrasound images at a reasonably high quality. Additionally, POCUS is more affordable and portable than traditional ultrasound machines. However, POCUS does not come with Doppler capabilities; clinicians are unable to make AS diagnoses using it. Recent advances in machine learning in the medical imaging community have shown deep learning to be successful in several ultrasound domains. The combination of deep learning with POCUS for AS diagnoses would allow for easily accessible screening.

1.2 Motivation

Timely diagnosis of AS is critical for patient survival. POCUS is a tool that enables widespread ultrasound availability but lacks the Doppler spectral capabilities needed for AS diagnosis. From a technical standpoint, the anatomical evaluation of AS is determined through two standard-plane echo views, the parasternal long-axis (PLAX) and parasternal short-axis AV level (PSAX). These two views, displayed in Figure 2 allow the AV to be visible from two different angles, and can provide information on the degree of calcification, and speed and range of motion, which are subsequently measured using Doppler. As shown in the figure, when AS becomes more severe, the opening of the AV is narrower and shows some visible signs of calcification [10]. A machine learning algorithm should be able to register the changes in AV in the echo and subsequently classify the AS severity. However, this is a difficult task as the algorithm needs to understand the degree of calcification and thickness of the AV, as well as how it affects the mobility of the cardiac cycle. Previous methods of automated AS assessment from ultrasound typically focus on just observing a single image from an echo, but the nature of AS echos indicates that temporal information to observe the opening and closing of the AV may be helpful as well. One of the most common uses for image-based deep neural networks is classification. Diagnosing AS severity can also be thought of as a classification problem, with network input being an echo cine. The resulting network output is the severity of

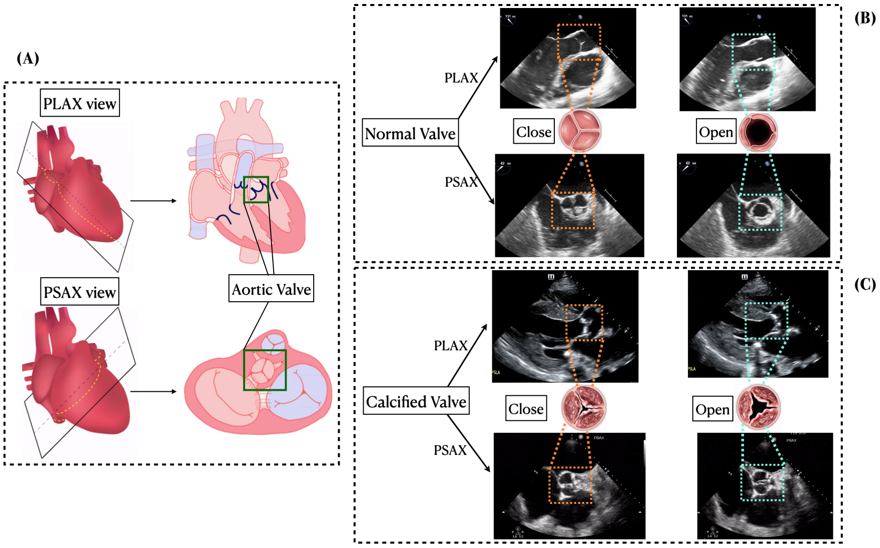

Figure 2: Parasternal long-axis and parasternal short-axis AV level views of aortic valve

the AS diagnosis (no AS, mild, moderate, or severe). In addition, deep neural networks can also successfully solve regression problems - where the network predicts a single value based on its inputs, rather than a class. AS diagnosis can also be approached as a regression problem, where the network tries to predict the numerical values of the three Doppler factors. However, from a practical standpoint, only one of the three measurements, Aortic Valve Area (AVA) is visually predictable. Therefore, a regression-based approach to this problem would be to use the neural network to predict the value of the AVA and then subsequently make a diagnosis based on the AVA value.

When clinicians make an AS diagnosis, the Doppler measurements are the main factors they use to determine the severity. However, the clinician also obtains other metadata information about the patient and their health history. This additional information consists of factors like the patient's age, weight, and other comorbidities. These data give the clinician a better understanding of the patient, which helps get a more accurate diagnosis. Current literature in the field of neural network-based AS diagnosis only uses echo data and fails to take into consideration any additional patient information that may be helpful. Based on these motivations and inspiration from previous works in the field, we present this paper detailing a machine-learning framework for AS severity diagnosis based on echo cine data. The main contributions and investigations in the paper are as three folds:

- Development of a deep neural network that classifies AS severity from echo videos.
- Exploration of classification networks versus networks based on regression of AVA.
- Integration of patient metadata into neural networks to assist with severity prediction along with echo.

1.3 Related Works

Recent developments in the field of deep learning serve as the baseline and backbone of the presented work. Specifically, the fields of automated AS image and video analysis and image-metadata fusion have been sources of influence for this work. In the field of image analysis for AS detection, Kang et al.[5] successfully classify severe AS from non-severe using a classifier based on computed tomography AV calcium scoring (CT-AVC) to a high level of accuracy. Chang et al.[3] use deep learning to segment calcified regions from CT images, and subsequently classify AS. Huang et al.[7] utilize a WideResNet to classify AS categories based on single cross-sectional echo images, then aggregate the predictions across the echo to form a patient-level prediction. They achieve an average prediction accuracy of 64.64%. Video analysis through deep learning has also been explored in the medical imaging field. Roshanitabrizi et al.[11] use Doppler measurement videos in the PLAX and PSAX views to detect rheumatic heart disease (RHD), which is another important aortic valve disease. Vimalesvaran et al.[16] use a random-forest-based classifier to detect the presence of AS in MRI cine

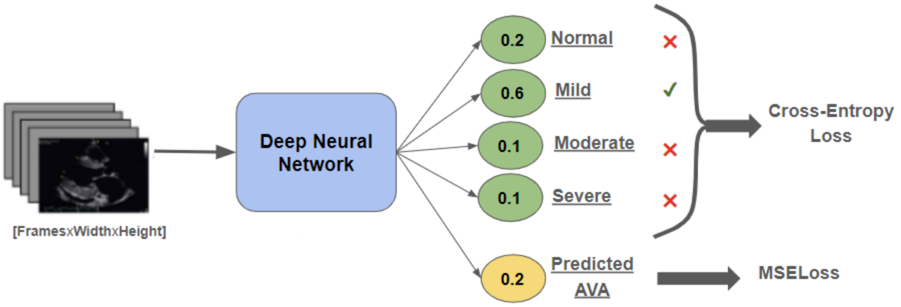

Figure 3: Abstract view of our network architecture. The green sections depicate the classification task, whereas the yellow one shows AVA registration task.

images. Another field of research that is relevant to our work is that of image-metadata fusion. This area involves combining separate non-image features along with relevant images. Minetto et al.[9] developed Hydra, a multi-head network that classifies geospatial land images, and fuses metadata with the inputs before the final prediction. This work was built on by Diao et al.[4], who use a MetaFormer consisting of transformer[15] combined with neural network layers to create a network that accepts both image and tabular data. They demonstrate their network performance on six open source datasets, improving upon image-only classification.

2 Techniques

2.1 Method Overview

Echo ultrasound videos are fed as training input to the deep neural model (Figure 3). The model improves its performance by adjusting its parameters via optimizing the network loss. The loss of a network is typically a metric that indicates how accurate the network predictions are and is selected based on the problem definition and network architecture. As shown in Figure 3, the model tries to predict each cine into one of four classes: normal (no AS), mild, moderate, or severe. After training, we evaluate the model on a separate testing set of echo cine videos. To validate the success of this classification, we use accuracy, which is the percentage of correctly classified patients. Three different setups were investigated with neural networks:

AS severity prediction based on classification: The output of the neural network is the probabilities of the cine belonging to a certain AS severity class, with the final prediction being the class with the highest probability. We use a softmax function (Eqn1) to determine the final predicted class. To train this model, we utilize cross-entropy[17] (Eqn2) as the loss function, which indicates how much the predicted probability diverges from the given label.

$$\sigma(p_i) = \frac{e^{p_i}}{\sum_{j=1}^{M} e^{p_j}} \quad for\ i = 1, 2, \ldots, M \tag{1}$$

$$Cross\ Entropy = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{2}$$

Where M is number of classes, log is the natural log, y is the binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is predicted probability observation o belonging to class c.

AS severity prediction based on AVA regression: The output of the neural network is a value corresponding to the predicted AVA. The predicted AVA value is then mapped to an AS severity class based on Table 1. We use mean squared error (MSE) as the regression loss function (Eqn3), which indicates the squared difference between the given value and the predicted value of AVA.

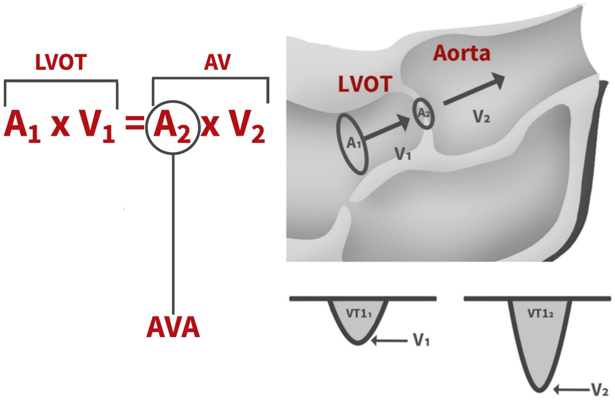

Figure 4: Abstract view of the network architecture. The green sections depicate the classification task, whereas the yellow one shows AVA registration task.

$$Mean\ Squared\ Error\ (MSE) = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{3}$$

Where D is the size of dataset, x_i is the predicted value of AVA for sample i, and y_i is the correct value of AVA for sample i.

For the regression task, we try to estimate AVA because it is the most important Doppler factor for determining AS severity. Also, it can be predicted visually, unlike Peak Velocity and Mean Gradient which is measurable only by Doppler. Figure 4 shows AVA and the continuity equation. The continuity equation states that the total blood volume that exits from the Left Ventricular Outflow Tract (LVOT) is equal to the total blood volume that enters the Aortic Valve (AV). In the clinical approaches, AVA is calculated using the continuity equation, as V_1 and V_2, which are the blood flow velocity, are captured using Doppler technology, and A_1 is easily calculated since LVOT is mostly stationary. Having the values of A_1, V_1, and V_2, clinicians can calculate A_2 which is AVA. In this project, however, we try to estimate AVA directly from the echo videos which is challenging since A_2 is changing between different frames due to the constant movement of the aorta.

AS severity prediction based on the combination of classification and AVA regression: The network outputs five values, four of which are probabilities corresponding to the AS classes and the other one is the predicted AVA value. The predicted AS category is determined from the classification probabilities similar to the first approach. The final loss function however, is a summation of cross-entropy loss over classes probabilities and MSE over the predicted AVA values.

$$Loss\ Function = Cross\ Entropy\ Loss + \lambda \times MSE \tag{4}$$

Where λ is a weighting hyperparameter which is assigned to 1 in the final experiments of this project.

2.2 ResNet-18 with (2+1)D Convolutions

As a baseline to our work, we propose a ResNet model with (2+1)D convolutions. This model takes echo videos as input and predicts AS severity classes under the mentioned setups in the method overview section. R(2+1)D[13] is a model based on the conventional ResNet[6] architecture, which is a stack of convolution layers on top of each other with residual connections. However, instead of using 3D Convolutions[14], it uses 2D convolution followed by a 1D convolution, decomposing spatial and temporal modeling into two separate steps. The model replaces the 3D convolutional filters of size $f \times w \times h$ with a (2+1)D block consisting of 2D convolutional filters of size $1 \times w \times h$ and temporal convolutional filters of size $f \times 1 \times 1$, where f is the number of the frames and w and h are frame dimensions. Compared to full 3D convolution, this (2+1)D decomposition offers two advantages. First, despite not changing the number of parameters, it doubles the number of

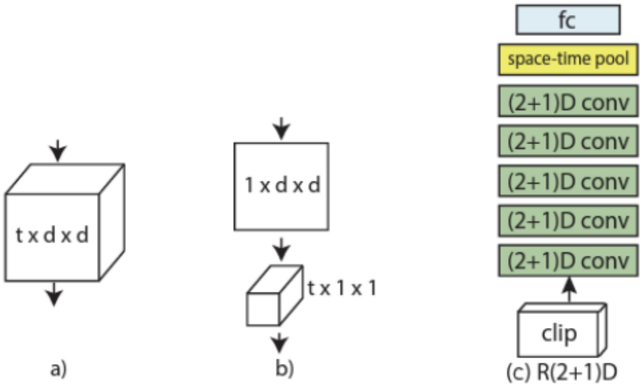

Figure 5: (2+1)D vs 3D convolution. The illustration is given for the simplified setting where the input consists of a spatiotemporal volume with a single feature channel. (a) Full 3D convolution is carried out using a filter of size t×d×d where t denotes the temporal extent and d is the spatial width and height. (b) A (2+1)D convolutional block splits the computation into a spatial 2D convolution followed by a temporal 1D convolution. (c) R(2+1)D are ResNets with (2+1)D convolutions. For interpretability, residual connections are omitted.

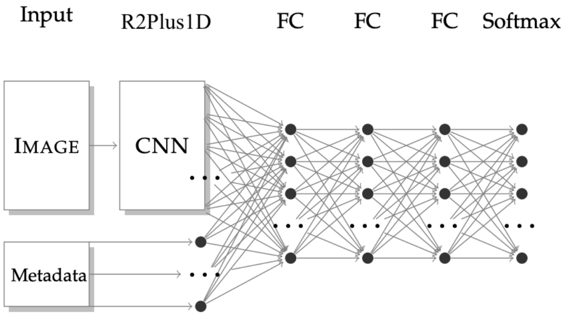

Figure 6: Architecture of MetaNet

nonlinearities in the network due to the additional ReLU[1] between the 2D and 1D convolution in each block. Increasing the number of nonlinearities increases the complexity of functions that can be represented. The second benefit is that forcing the 3D convolution into separate spatial and temporal components renders the optimization easier. This is manifested in lower training error compared to 3D convolutional networks of the same capacity. Figure 5 shows the comparison between 3D and (2+1)D convolutions. Finally, we stack 18 (2+1)D convolutional layers and apply residual connections between them to build the final model. The architecture of this model is shown in Figure5. In the rest of the paper, we refer to this model as **R2Plus1D** as an abbreviation.

2.3 MetaNet

MetaNet (Figure 6) is based on the architecture behind Hydra [9] and built on top of the R2Plus1D network. Typically in R2Plus1D, the final fully connected layer would result in the classification output. In the case of MetaNet, this final fully connected layer is removed. Instead, the patient metadata is normalized and then appended to the embedding of the cine. The combined embedding and metadata are then fed into three fully connected layers, then a classification category is outputted. Having three fully connected layers after appending the metadata allows for deeper relationships and nonlinearities between the metadata, which could result in better performance than just having one fully connected output layer after appending the metadata.

Table 2: Quantitative Results for AS Prediction

Model	Regression		Classification		Regression & Classification	
	mean	std	mean	std	mean	std
R2Plus1D	22.19%	0.00%	65.29%	1.13%	70.59%	4.16%
MetaNet	22.19%	0.00%	65.63%	2.81%	71.82%	3.07%

3 Results

3.1 Dataset

The data consists of 9115 echos gathered from 2247 patients from Vancouver General Hospital (VGH). Each patient has a set of metadata and multiple echo videos in both the PLAX and PSAX views. These views allow us to see the heart from different angles. These ultrasound were labeled by clinicians into 4 categories, which are Normal (no AS), Mild, Moderate and Severe. In addition, each echocardiogram has additional metadata associated including age, other risk factors, hospital, and date. The dataset consists of 5055 PLAX echos and 4060 PSAX echos. Moreover, it has 3600 Normal, 2100 Mild, 1700 Moderate, and 1715 Severe samples. We randomly split our data into training, validation, and testing, with 7000, 1115, and 1000 samples respectively.

3.2 Training Details

For each network, each experiment was run three times for 40 epochs. After each epoch, validation accuracy and loss are computed using a separate validation set. For training, a batch size of four was used, as well as a decaying learning rate. In addition, the number of normal patients in the dataset was much higher than in all the other classes. To prevent this from negatively affecting our network to favor a certain class, a class-balanced sampler was used. The class-balanced sampler weighted the samples in each batch based on how frequently their AS class appeared in the training data. When looking at the metadata information, we noticed that some of the features may not necessarily be directly related to AS severity. The presence of such features in our data would thus cause the model to overfit. These irrelevant features consisted of information like which hospital the patient was at, the date they were admitted, patient ID at the hospital, and any other information that is not relevant to AS severity. These features were dropped from the metadata. The final metadata features that were used for the model were: rhythm, risk factors (smoking, diabetes, high blood pressure, etc.), right ventricular function, mitral valve function, aorta information, and left ventricle grade.

3.3 Results

For both R2Plus1D and MetaNet, the regression experiment of predicting the AVA value achieved a mean testing accuracy of 22.19%. This is equivalent to always predicting a singular class, and showing no sign of successful learning. This can be attributed to the fact that it is difficult to determine AVA without Doppler measurements. In addition, AS diagnosis is made using AVA in combination with other factors, not just AVA, which could also contribute to poor performance when making the prediction just based on one factor. R2Plus1D and MetaNet achieved testing accuracies of 65.29% and 65.63% respectively for the classification task. This shows that the model was able to learn some valuable information, as it is much higher than the accuracy of randomly guessing or guessing all one class. The combination of classification and regression for prediction achieved the best results. R2Plus1D achieved an accuracy of 70.59%, while MetaNet achieved an accuracy of 71.82%, which improves upon the performance of both the regression and classification alone. This suggests that combining the methodologies and loss functions allows the network to learn more information from the cine echos than any given individual task. The quantitative results are shown in Table 2. Moreover, as we can see in Figure 8, the R2Plus1D model overfits on the training data and the validation loss starts to increase after 15 epochs. On the other hand when using MetaNet, as shown in Figure 7, training loss and validation loss decrease continuously. This shows that using metadata helps with the generalization of the model. More charts are provided in the Appendix section for comparison between training and validation of the models.

Table 3: Ablation Study on Number of AS Classes

AS Classes	Accuracy	
	Training	Testing
4-Classes	91%	70%
3-Classes	93%	78%
2-Classes	95%	89%

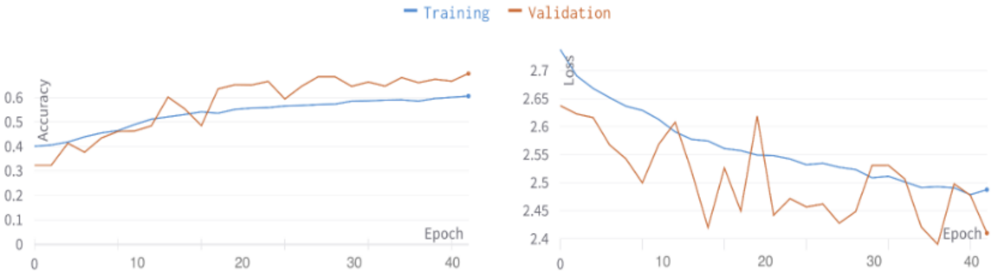

Figure 7: Training and Validation charts for MetaNet model under classification & registration tasks. The left chart shows the accuracy and the right chart shows loss over epochs.

When looking at the predictions generated by the models, it appeared that the models struggled to predict the Moderate AS class. Moderate AS cases were almost always predicted as either severe or mild. Based on this information, we conducted an ablation study on the number of AS classes (Table 3). The number of classes was reduced from four to three, combining the mild and moderate AS classes. This resulted in an 8% increase in testing accuracy. When the number of classes was further reduced to only two classes, Normal (no AS) and AS, the testing accuracy improved to 89%, which indicates that our model successfully learns indicators of AS.

4 Conclusions

In this project, we introduced an automated end-to-end framework for AS severity diagnosis based on a deep learning model. We performed three experiments, classification, regression, and a combination of classification and regression, on two different models, R2Plus1D and MetaNet. We demonstrated that the classification and classification plus regression tasks showed significant signs of being able to

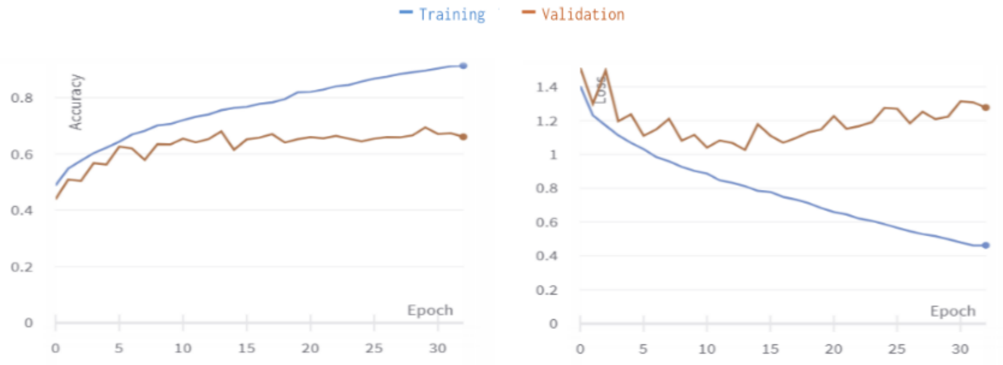

Figure 8: Training and Validation charts for R2Plus1D model under classification & registration tasks. The left chart shows the accuracy and the right chart shows loss over epochs. Blue is for training set and orange is for validation set.

learn the different AS classes. In addition, we conducted ablation studies that showed that reducing the number of classes greatly improved the accuracy of the models.

4.1 Discussion

When comparing the R2Plus1D model against MetaNet, there is no significant improvement from adding the additional metadata. Accuracies across the three experiments were all relatively similar. However, the metadata did not cause the model to perform much worse, which indicates that the metadata information is still relevant in making the predictions. With more work and regularization around the selection of important metadata features, it is quite possible that the MetaNet model can keep improving. Nevertheless, one advantage associated with using the metadata is that there is relatively less overfitting when compared to the R2Plus1D model. While there are techniques to avoid overfitting, it seems that adding the metadata provided the models with enough new data to avoid this problem.

Both models struggled greatly with the standalone regression task, consistently converging to a single bad value. Predicting the AVA alone is very difficult, especially just based on an echo. There was not enough visual or metadata information to allow the network to learn the AVA values. In addition, clinicians use all three Doppler factors when making a diagnosis, so giving the model only one of the three may also result in poor performance.

Based on the experiments, the combination of classification and AVA produced the best results. While the regression task proved unsuccessful alone, it ultimately still provided some valuable information, just not enough for the model to learn without additional help. When combined with the classification task, it was able to help generate more accurate predictions.

4.2 Future works

For future work, there are many ways this framework can be improved upon and extended. One such possible extension is to provide uncertainty estimates with the predictions. This would indicate how confident the model is with its predicted severity class. In addition, receiving a high uncertainty estimate is more informative than simply guessing a random class. In a clinical setting, a high uncertainty estimate would notify a clinician that the model is not able to make an accurate prediction, and assistance from Doppler measurements and a clinician is needed.

Another improvement on this project would be to add explainability and interpretability to the model. Upon generating a prediction, the model would then indicate which frames helped decide the prediction, or provide a segmentation of which section of the echo was important. This would help further AS research and help improve echos as a whole as ultrasound technicians would get a better idea of what they should be scanning.

Furthermore, this project could be extended to give patient-level instead of video-level predictions. Some patients in the dataset have multiple echo videos corresponding to them. Currently, our model generates AS predictions on a per-video basis. However, it may be helpful for the model to leverage multiple videos of the same patient, and thus generate a predicted AS severity class for each patient rather than each video.

References

[1] Abien Fred Agarap. “Deep Learning using Rectified Linear Units (ReLU)”. In: *CoRR* abs/1803.08375 (2018). arXiv: `1803.08375`. URL: `http://arxiv.org/abs/1803.08375`.

[2] Blase A Carabello. “Introduction to aortic stenosis”. In: *Circulation research* 113.2 (2013), pp. 179–185.

[3] Suyon Chang et al. “Development of a deep learning-based algorithm for the automatic detection and quantification of aortic valve calcium”. In: *European Journal of Radiology* 137 (2021), p. 109582.

[4] Qishuai Diao et al. “MetaFormer: A Unified Meta Framework for Fine-Grained Recognition”. In: *arXiv preprint arXiv:2203.02751* (2022).

[5] Nam Gyu Kang et al. “Performance of prediction models for diagnosing severe aortic stenosis based on aortic valve calcium on cardiac computed tomography: incorporation of radionics and machine learning”. In: *Korean journal of radiology* 22.3 (2021), p. 334.

[6] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: `1512.03385`. URL: `http://arxiv.org/abs/1512.03385`.

[7] Zhe Huang et al. “A New Semi-supervised Learning Benchmark for Classifying View and Diagnosing Aortic Stenosis from Echocardiograms”. In: *Proceedings of the 6th Machine Learning for Healthcare Conference (MLHC)*. 2021. URL: `https://tmed.cs.tufts.edu/papers/HuangEtAl_MLHC_2021.pdf`.

[8] Writing Committee Members 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”. In: *American College of Cardiology Foundation Washington DC* 77.4 (2021), e25–e197.

[9] Rodrigo Minetto, Maurıcio Pamplona Segundo, and Sudeep Sarkar. “Hydra: an Ensemble of Convolutional Neural Networks for Geospatial Land Classification”. In: *CoRR* abs/1802.03518 (2018). arXiv: `1802.03518`. URL: `http://arxiv.org/abs/1802.03518`.

[10] Liam Ring et al. “Echocardiographic assessment of aortic stenosis: a practical guideline from the British Society of Echocardiography”. In: *Echo Research and Practice* 8.1 (2021), G19–G59.

[11] Pooneh Roshanitabrizi et al. “Ensembled Prediction of Rheumatic Heart Disease from Ungated Doppler Echocardiography Acquired in Low-Resource Settings”. In: (2022), pp. 602–612.

[12] Geoff Strange et al. “Poor long-term survival in patients with moderate aortic stenosis”. In: *Journal of the American College of Cardiology* 74.15 (2019), pp. 1851–1863.

[13] Du Tran et al. “A Closer Look at Spatiotemporal Convolutions for Action Recognition”. In: *CoRR* abs/1711.11248 (2017). arXiv: `1711.11248`. URL: `http://arxiv.org/abs/1711.11248`.

[14] Du Tran et al. “C3D: Generic Features for Video Analysis”. In: *CoRR* abs/1412.0767 (2014). arXiv: `1412.0767`. URL: `http://arxiv.org/abs/1412.0767`.

[15] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: `1706.03762`. URL: `http://arxiv.org/abs/1706.03762`.

[16] Kavitha Vimalesvaran et al. “Detecting Aortic Valve Pathology from the 3-Chamber Cine Cardiac MRI View”. In: (2022), pp. 571–580.

[17] Zhilu Zhang and Mert R. Sabuncu. “Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels”. In: *CoRR* abs/1805.07836 (2018). arXiv: `1805.07836`. URL: `http://arxiv.org/abs/1805.07836`.

Appendix

1 Code

Our code for this project is publicly available at https://github.com/Armin-Saadat/AS_project

2 Professionalism and Ethics

2.1 Ethical

Equity, diversity, and inclusion

Equity, diversity, and inclusion are important parts of any engineering project. Firstly, this issue must be addressed by the research team. Diverse teams consist not only of different areas of expertise but also people from different intersectional backgrounds. Ensuring the research team is diverse also involves making sure that everyone on the research team is treated fairly and given equal opportunities. This will not only result in better research due to the diversity but also a more equitable work environment. In addition, this study involves patient data used for network training. While information about the patient is not given with the data, undoubtedly different biological sexes and races will have some biological differences. To ensure that the model can work fairly in clinical cases, it is important to ensure that the training data comes from a diverse distribution of patients.

2.2 Regulatory

Regulations, codes, bylaws, and standards

This study involved using patient data collected from hospitals. Using this data, it is important to take into account regulations and standards around protecting the information and privacy of the patient. No personal or identifiable details should about the patient be stored with the data. In addition, there are many bylaws in Canada surrounding participation in research studies, highlighting the importance of patient consent, which would also need to be taken into consideration for this project. All data would need to be stored in a secure share; no echos should be stored on any researcher's personal computer. This ensures better security of the data and also prevents researchers who no longer work on the project or team from accessing the data.

2.3 Technical

Technical regulations, codes, and standards

In machine learning projects, researchers need adequate computation power to conduct experiments and train deep neural networks. For visual medical imaging problems like ours, GPUs are the computation units that provide the necessary resources. In this project, we used our laboratory's GPU shared with fifteen other researchers. The demand for GPU is usually higher than its supply, so there are some regulations about booking and using it. In this project, we prepared a detailed schedule in which we specified which experiments we wanted to run, what GPU power we needed, and when to conduct our experiments. Therefore, we managed to book the required GPUs and utilize them according to the standards of the lab. We completed our experiments successfully while complying with our share of the GPUs. The other technical standard in programming projects is maintaining the source code's integrity and readability, which requires proficiency in clean coding and excellent communication between programmers and teammates. To this end, we used GitHub as the project version control, which helped us keep the code clean, usable, and professional.

2.4 Communications and Leadership

Consulting/business skills

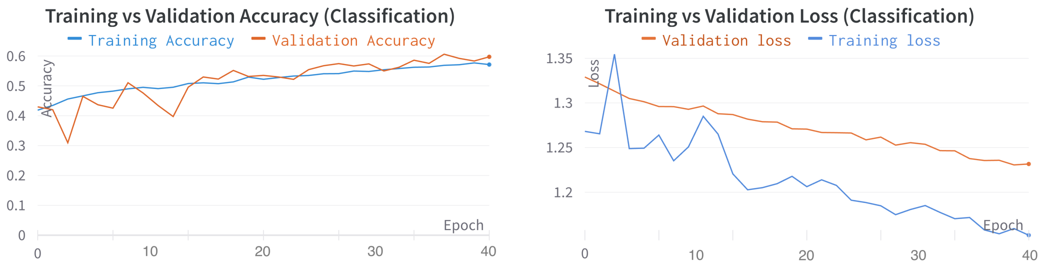

Figure 1: Training and Validation charts for MetaNet model under classification task. The left chart shows the accuracy and the right chart shows loss over epochs.

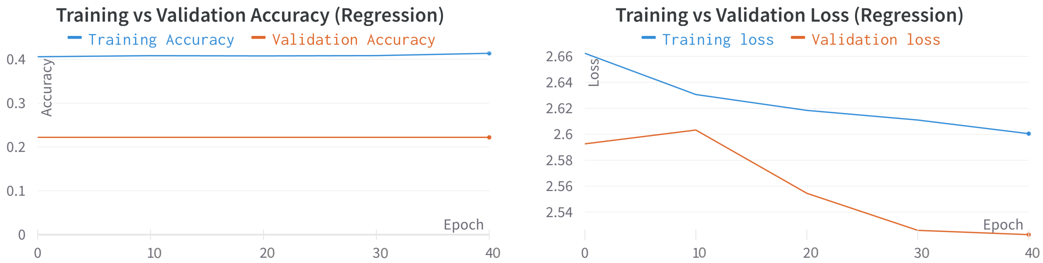

Figure 2: Training and Validation charts for MetaNet model under AVA regression task. The left chart shows the accuracy and the right chart shows loss over epochs.

Consulting with experts is the key to escaping dead ends and progressing in the project. In our work, after designing and training the machine-learning models, we realized that the accuracy of the model is particularly poor in one class of echo videos. The model was performing acceptable in general but, was performing poorly on the samples with Moderate aortic valve stenosis. Hopefully, our research team has some connections to clinicians as we mainly work on medical problems. We askes two clinicians to comment on the situation, and it turned out that detecting and classifying the Moderate cases is difficult even for experts. Therefore, it became reasonable that out model struggled to handle patients with moderate AS. Learning this information, we changed our strategy and considered a margin of error for these cases. In other words, we focused on classifying AS severity into two classes(normal or with AS). The clinicians also commented on some samples from other classes as well. We realized that our model struggles when facing out-of-distribution data, which is something we plan to solve in future works.

3 Additional Figures

Additional figures of the training and validation accuracy and loss can be found here. Figure 1 described the progress of the classification task, while Figure 2 shows the regression task.