Ejection Fraction Prediction from Echocardiograms

Vasu Kaker, Daniel Chung, Mindy Somin Lee, Irbaz Riaz,
Yongyi Zhao, and Sudheesha Perera

Abstract

Heart failure affects approximately 6.2 million people in the US, and in 2018 alone was mentioned on over 350,000 death certificates [1]. Moreover, it accounts for $30.7 billion of healthcare expenditures on an annual basis [2]. Reliable and expedient diagnosis of heart failure when symptoms first present is critically important to maintaining quality of care and preventing suffering, and thus there is growing interest in the efficacy and safety of automated models that can provide such assessments without expert intervention. In this work we take state-of-the-art deep learning models for Video Action Recognition and apply them to the binary classification of echocardiogram heartbeats int Healthy (Ejection Fraction above 40%) and Unhealthy (Ejection Fraction below 40%) patients. The Swin Transformer achieves good performance on the EchoNet-Dynamic Dataset, achieving a test AUC score of 0.88 and a test accuracy of 90% on this binary classification task. We also train an R3D model on the same binary classification task, achieving a test AUC of 0.92 and a test accuracy of 92%, exceeding the performance of previous comparable studies with only 30% of the training set [3].

1. Introduction

Heart failure is the chronic impairment of the heart's ability to effectively pump blood, resulting in a constellation of signs and symptoms that amount to a marked reduction in patients' overall health and quality of life. Heart failure affects approximately 6.2 million people in the US and in 2018 alone was mentioned on over 350,000 death certificates [1]. Moreover, it accounts for $ 30.7 billion of healthcare expenditures on an annual basis [2]. However, there exist a range of effective interventions – both medication-based and procedural – that have been shown to prolong life and reduce morbidity in cases where a heart failure exacerbation can be detected in a timely manner [4]. Thus, from a clinical perspective, reliable and expedient diagnosis of heart failure is critically important to maintaining quality of care and preventing downstream sequelae of unaddressed heart failure exacerbations.

In addition to the careful assessment of symptoms and physical exam findings, the interpretation of imaging is a key component in the diagnosis of heart failure. In particular, the use of echocardiography, i.e. ultrasound of the heart, allows for the assessment of left ventricular ejection fraction (EF), a key metric in establishing the diagnosis and prognosis of heart failure. EF represents the fraction of blood that exits the left ventricle during the systolic phase of the cardiac cycle. Cases of heart failure in which the ejection fraction is below 40% are referred to as heart failure with reduced ejection fraction (HFrEF).

Of particular interest are automated approaches to interpretation of heart failure. Given that interpretation of echocardiography is typically performed manually by a trained clinician, physician bias and lack of standardization are evident issues. Additionally, lack of trained cardiologists in rural or low resource settings means that patients see delay of reading or lack of evaluation, leading to downstream negative health effects [2]. Therefore, if a tool for automated EF prediction were to exist, the implications would be massive for the early detection, verification, and treatment of heart failure [5].

From a machine learning perspective, addressing this problem is challenging. Echocardiograms are stored as video files that contain a certain degree of variability, and suffer issues of video quality and lack of standardization. This is paired with all of the standard challenges of handling video data in performing a prediction task [6] . Furthermore, it is well documented the inherent challenges of inferring cardiac function from echocardiography in general, given the limited nature of the two-dimensional imaging modality and the true complexity of cardiac physiology and kinesiology [7]. Others have tried a slew of different approaches to a multitude of cardiac pathologies, from video transforms to detect structural heart defects to utilizing CNN-LSTMs to predict cardiomyopathy and cardiac amyloidosis [8-11].

In this work, we analyzed 10,036 echocardiograms from the EchoNet dataset first introduced by Ouyang et al. to predict low EF (defined as EF below 40%) utilizing two best-in-breed video transfers [5]. The EchoNet dataset represents the largest labeled medical video dataset made available publicly to researchers and medical professionals, and our analysis represents an assessment of the ability of a novel transformer approach to improve existing tools for automated EF determination.

Generalizable Insights about Machine Learning in the Context of Healthcare

At a high level, this work is offered as evidence of the efficacy of standard, best-in-breed machine learning techniques in solving a critically relevant clinical prediction task. In particular, we aim to display that a transformer-based approach to predicting low EF from raw echo reads is efficacious, with associated predictive capability at or above what would reasonably be required for real clinical practice. Ultimately, accurate assessment of ventricular function is critical for patients with heart failure and associated conditions. Therefore, our models serve as an encouraging stepping stone towards developing robust clinical decision support tools that would aid in assessment of patients with suspected heart failure. Potential clinical applications include: (1) automated assessment of ventricular function in low resource settings where trained cardiologists are not available, (2) expedited, actionable insight in high resource hospital settings where cardiology input is not immediately available

2. Related Work

The EchoNet Dynamic dataset was first introduced by Ouyang et al. along with the performance of three 3D convolutional architectures for video classification used to assess ejection fraction as a numerical outcome (ie. percentage). The best of their models, an R3D convolutional neural network built on ResNet-18, was able to achieve an R-squared of 0.71 when assessed against ground truth ejection fraction [5]. Previous works by Asch et al had displayed R-squared of 0.95 on their proprietary database of roughly 50,000 echocardiograms [7]. The authors later followed this work with a Nature paper in 2020 in which they presented results from a model specifically trained to predict cases of low ejection from, this time displaying an AUC of 0.97 [6]. More recently in 2022 Almadani et al. utilized recognition (VAR) neural networks to perform binary classification of a far more complex set of echocardiograms, and reported accuracy of 90.17% with inference time as low as 25.11 seconds using a Gate Shift Network with BNInception as its backbone [3]. Outside of binary classification of low ejection fraction, as well as numerical prediction, a number of associated models for determining other aspects of heart function based on echocardiograms have also been introduced into the literature. Notably, Zhang et al. (2018) were able to detect hypertrophic cardiomyopathy, cardiac amyloidosis, and pulmonary arterial hypertension with C statistics of 0.93, 0.87, and 0.85, respectively, as well as determine left ventricular mass, left ventricular diastolic volume, and left atrial volume within narrow range of machine-derived values [9].

More recently, general video transformers have been applied to various prediction tasks using echocardiograms. For example, Jafaeezadeh et al utilized a deep learning model with inception architecture as the backbone to 71% accuracy in the task of detecting mitral valve dysfunction, providing preliminary evidence that deep learning and transformers in echocardiographic videos can render quick, precise, and stable evaluations of various cardiac pathologies. [10]. Dai et. al applied a novel Cyclical Self-Supervision (CSS) method for learning video-based LV segmentation, and showed that their method outperformed alternative semi-supervised methods to achieve MAE of 4.17, which is competitive with state-of-the-art supervised performance, using half the number of labels [11]. Apart from general video transformers and the convolutional neural networks presented above, CNN-LSTM models have also been applied in this context. Most recently, Hwang et al achieved an accuracy of 92.8% on detection of LVH (i.e. hypertensive heart disease [HHD], hypertrophic cardiomyopathy [HCM], and light-chain cardiac amyloidosis [ALCA] [12]. To our understanding, this is the exploration of utilizing two state of the art video transformers approaches (SwinVideoTransformer and R3D) to assess the ability of these models to predict low ejection fraction from echocardiograms alone.

3. Methods

3.1. Swin3D Transformer Model

The Video Swin Transformer was introduced in a 2021 paper by Microsoft Research [13]. The model takes substantial inspiration from the original Transformer, introduced in 2017. The model was amongst the top performers of models of similar sizes on relevant classification benchmarking datasets such as Kinetics-400. Larger models, albeit better performing on this dataset, may be unsuitable for our purposes [14].

Similar to the original transformer, the model relies on using attention mechanisms on the video token sequence. For an input video of dimension TxHxWx3, it creates a total of (T/2) * (H/4) * (W*4) tokens of dimension 2*4*4*3 (frames, height, width, channel) tokens each of dimension 96, further projecting these tokens to an arbitrary dimension C. Then, in non-overlapping windows of size P*M*M corresponding to the relevant tokens in that window, the model applies local self-attention, creating a total of (T/P) * (H/M) * (W/M) non overlapping 3D windows each of size [P * M * M]. The attention function, as defined in the original Swin paper, is given explicitly below [13]:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V,$$

Local self-attention is then performed within these 3D windows, after which a combination of self and cross attention is performed between each non-overlapping 3D windows as

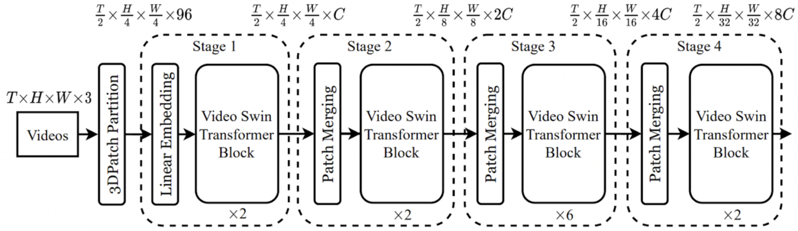

$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times 96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
Stage 1
Stage 2
Stage 3
Stage 4
$T\times H\times W\times 3$
Videos
3DPatch Partition
Linear Embedding
Video Swin Transformer Block
Patch Merging
Video Swin Transformer Block
Patch Merging
Video Swin Transformer Block
Patch Merging
Video Swin Transformer Block
×2
×2
×6
×2

Figure 1: Video Swin Transformer Architecture, Liu et al. 2022

illustrated in Figure 1. Similar to a typical Transformer architecture, these layers of self-attention are interspersed with nonlinearities and culminate in classification on the decoder side of the transformer. Furthermore, the video also uses positional bias in its Attention Mechanism (see term B) in order to improve its weighting on nearby tokens over ones that are farther away.

3.2. R3D Model

The R3D model was one of the best-performing approaches that Ouyang et al. (2019) evaluated when tackling the echocardiogram classification problem [5]. Proposed by Tran et al. (2018) as an approach for video learning [16], R3D is especially apt for video data like echocardiograms because it uses 3D convolutional filters. These not only span the 2 spatial dimensions of a frame but also the temporal dimension that emerges from stacking them, allowing it to learn patterns not just spatially within frames but temporally between them. R3D in particular uses spatiotemporal kernels of shape 3× 3 × 3.

3D CNN architectures are typically shallow, but R3D combines the advantages of spatiotemporal kernels for video learning with the advantages of deeper architectures for better overall performance. In particular, R3D builds on a ResNet34 architecture, which consists of a very deep CNN architecture with 34 layers. ResNet owes much of its performance to residual blocks that comprise its architecture: each residual block has a shortcut connection that allows a gradient to be directly backpropogated to earlier layers, which eases gradient calculations and cumulatively mitigates overfitting.

In short, R3Ds use spatiotemporal kernels to capture both spatial and temporal patterns in video data, and they rely on the depth of the ResNet34 architecture to give this spatiotemporal reach a powerful underlying learning architecture. They thus make a perfect fit for video classification.

3.3. S3D Model

The separable 3D CNN (S3D) model was introduced by Xie et al. (2018) as a deep learning approach that can learn video features without the use of spatiotemporal kernels [17]. Instead of using one kernel of shape kt × k × k, where kt is the temporal size and k is height and width, S3D uses a spatial 1 × k × k kernel followed by a temporal kernel kt × 1 × 1. It is called a "separable" 3D CNN because there are separate kernels to handle the temporal and spatial dimensions, yet it is 3D because these kernels combine to cover both

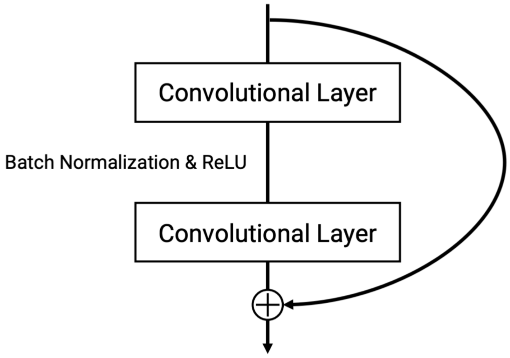


Figure 2: Example of a shortcut ("skip") connection within a residual block.

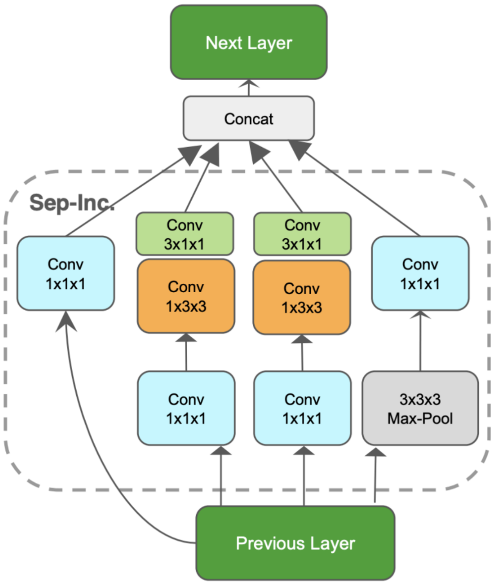


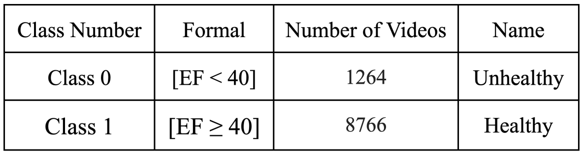
Figure 3: A 3D temporal separable inception block featuring paired spatial and temporal convolutional filters.

spatial dimensions and the temporal one in the end. This architecture reduces the number of learnable parameters compared to other video CNNs, making S3D a very attractive method for its computational efficiency. We are currently in the process of implementing this model on our dataset.

3.4. Training

Training was prohibitively expensive on local devices, mainly due to memory constraints. We initially approached this by downscaling all videos to a resolution of 56 × 56 instead of

Table 1: Cohort Split Based on Ejection Fraction (EF)



Class Number	Formal	Number of Videos	Name
Class 0	[EF < 40]	1264	Unhealthy
Class 1	[EF ≥ 40]	8766	Healthy

112 × 112, grayscaling all videos from 3 channels (RGB) to 1 (gray), and downsampling the videos such that we only considered 1 every 4 frames. However, the most effective solution was to utilize the MIT Satori cluster, which gave us the computational resources to train both the Swin3D transformer and the R3D model for 20 epochs each.

We used an Adam optimizer and achieved our best results based on validation sets using a learning rate of 1e-5 and a weight decay of 5e-4. To further manage our memory resources, we also used gradient accumulation, which involved accumulating gradients over successive mini-batches of size 20 before using them to update the model. This comes with a time cost, however, as there are more forward and backward passes involved with each parameter update. Nevertheless, because memory was our tighter constraint, we opted to move forward with gradient accumulation in our modeling process. We also chose gradient accumulation as opposed to batching since our videos were of varying time length.

4. Cohort

4.1. Cohort Selection

Our dataset consists of 10,030 echocardiograms. Ejection fraction (EF) served as our target variable, and we were interested in whether it was above or below 40%. This is because an EF below 40% warrants certain medical treatments that an EF above 40% does not. As with many health-related target variables, this caused a class imbalance program, as 87% of the EF were greater than or equal to the 40% cutoff and only 13% were below. The precise figures can be seen in Table 1.

The distribution of EF in our dataset was left-skewed with a mode around 60%, which is to be expected given that a normal value for EF is above 50% [4]. Specifically, the mean EF is 0.55, which is slightly lower, because there is a group of EFs between 20-50% that drag the average down. Notably, there were virtually no records where EF exceeded 80% or fell below 10%. The standard deviation is 12.

Importantly, we did not standardize frame length for each video, so the value of n changes for each record in our dataset since we didn't not want to remove valuable data to be fed into our models. An investigation of frame length distribution shows that our echocardiograms have 173 frames on average with a standard deviation of 47 frames. This distribution appears normal. Keeping a variable frame length was important because it reflects medical reality;: echocardiograms come in different lengths. Training a deep learning model on echocardiograms of frame length 170, for example, may not generalize well to echocardiograms of frame length 50 or 400. Non-standardized frame lengths thus functioned as a robustness feature.

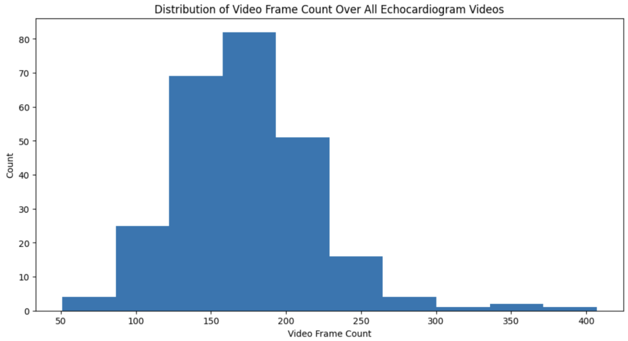


Figure 4: Distribution of Frame Length for All Echocardiograms

Our dataset split resulted in 7,458 training videos, 1,284 validation videos, and 1,277 test videos. Because of the underlying class imbalance, however, only 948 (13%) of the training videos were class 0. One risk of training on such imbalanced datasets is that the model can learn to maximize accuracy by always predicting the majority class. Indeed, our early modeling attempts found that models trained on this dataset succumbed to this result. To mitigate this risk, we deliberately undersampled videos of class 1 from the training set such that our new train data had 948 class 0 videos and 948 class 1 videos. This reduced our training set to 1,896 videos, but it made it perfectly balanced and less likely to produce a monopredictive model.

Downsampling of class 0 would prevent taking full advantage of the dataset and introduce the risk of leaving out important information included in the original dataset. Therefore, after we obtained access to the MIT Satori cluster, we also attempted another commonly used method in computer vision, i.e. data augmentation [18], in order to make full use of the dataset, prevent overfitting, and at the same time rebalance the two classes. To be clinically meaningful and comparable with the benchmark work [3], we applied transforms including ColorJitter (brightness, contrast, saturation, hue), Grayscale, combination of ColorJitter and Grayscale, Brightness, and Sharpness. Example of the data augmentation is shown in Figure 5. After data augmentation, class 1 sample size increased 5 times and became similar to the size of class 0. Due to time constraints, however, we were not able to include the augmented data in our training set; our training set finally consisted of 1,896 videos for the purpose of binary prediction. Our training set for numerical prediction of EF, however, consisted the full 7,458 videos as there is no such thing as class imbalance for exact quantity prediction.

5. Results

5.1. Swin3D for Binary Classification

We train a Swin3D Transformer model on our training dataset for 20 epochs. At each epoch we evaluate this same model on our validation set, saving the model parameters with the highest validation AUC score.

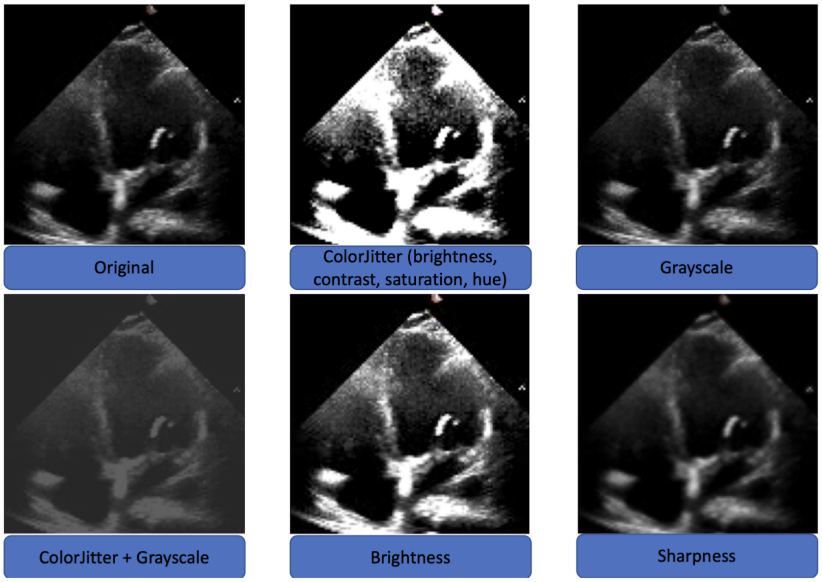
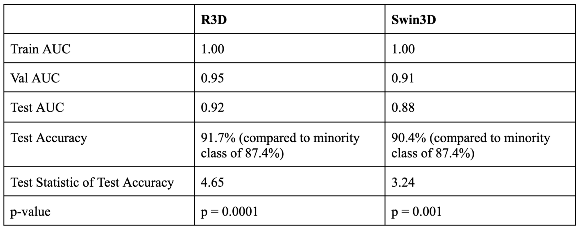


Figure 5: Examples of a few types of image augmentations when applied to the same original video frame.

Table 2: R3D Model Outperforms Swin3D



	R3D	**Swin3D**
Train AUC	1.00	1.00
Val AUC	0.95	0.91
Test AUC	0.92	0.88
Test Accuracy	91.7% (compared to minority class of 87.4%)	90.4% (compared to minority class of 87.4%)
Test Statistic of Test Accuracy	4.65	3.24
p-value	p = 0.0001	p = 0.001

For the purpose of assigning a label to each of our prediction probabilities, we determine the binary classification threshold that maximizes accuracy from the epoch with the highest validation AUC score.

We determine the statistical significance of our findings as follows. We first take the null hypothesis that our model simply predicts the majority class. The majority class consists 87.47% of the test set. We then consider that our model achieves an accuracy of 90.36% on our test set. Modeling both the null hypothesis and our model's predictions as a bernoulli variable with parameter theta = accuracy_score for a sample size of 1,277, we calculate our test statistic to be 1.71, and using a right-tailed test, our p value to be 0.04

Our train loss decreases, indicating model fit to data. However, our validation loss does not improve substantially, despite our AUC score increasing from 0.6 to 0.91 on the validation set over the course of 20 epochs.

We attribute this phenomenon to the possibility that our model is becoming more confident in all of its predictions – both correct and incorrect – over the course of training.

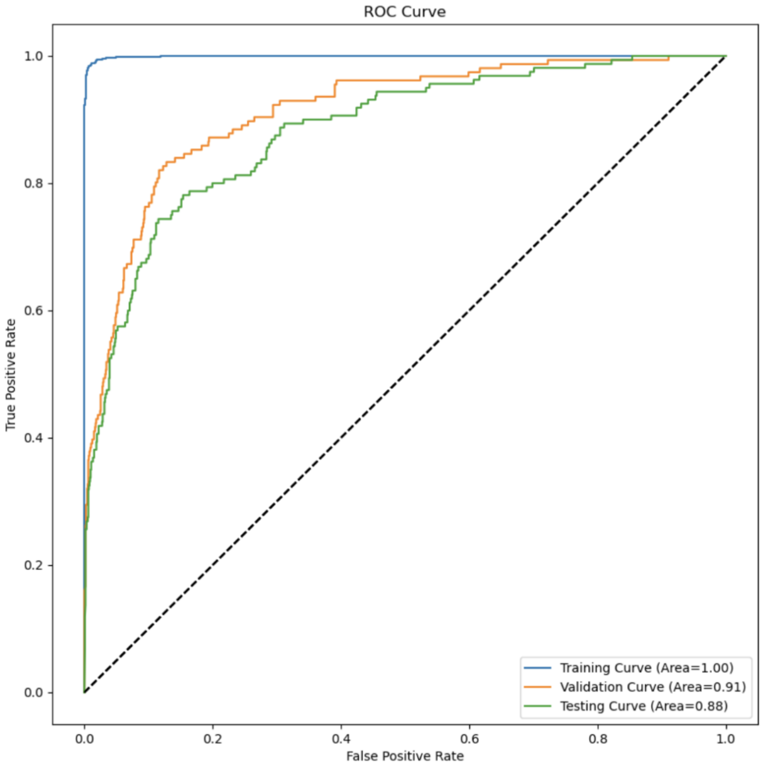


Figure 6: AUC Curve for Swin Model Performance

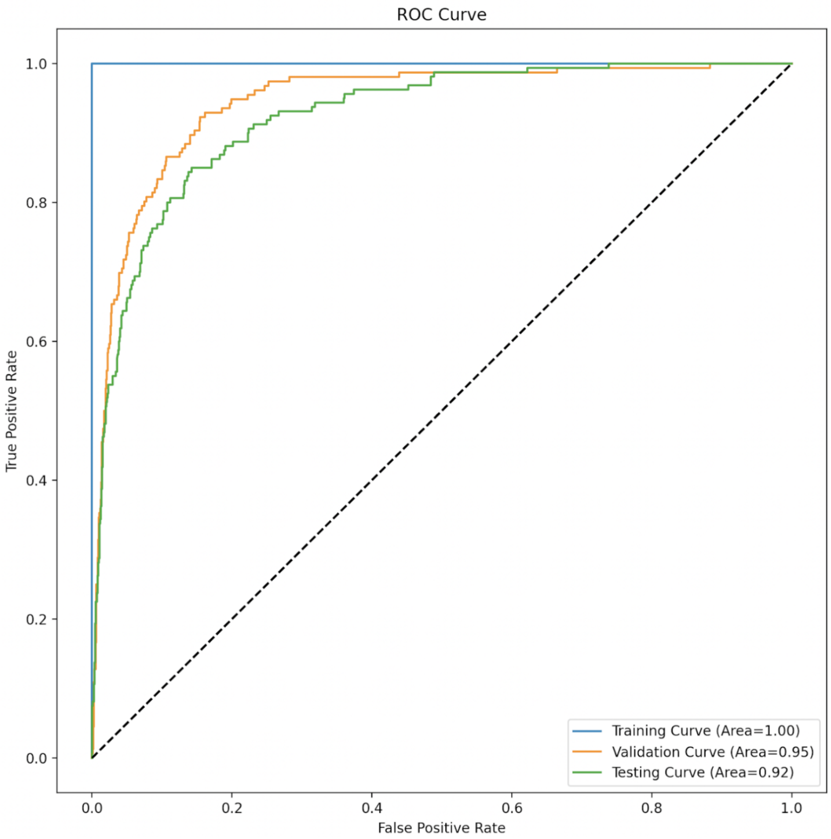


Figure 7: AUC Curve R3D Binary Classification

The improvements in loss in the correct predictions is offset by the decrease in our loss for incorrect predictions. The AUC score nevertheless increases as the model becomes more certain in its predictions and fits to the dataset.

5.2. R3D for Binary Classification

Using the R3D Model for binary classification we obtain excellent results. The high train AUC explains that the model has sufficient fit to the training set by the time it achieves the highest validation AUC score. The strong performance on unseen data is further demonstrated by its excellent Test AUC of 0.92 [Figure 7].

5.3. Swin3D for Numerical Prediction

We train the Swin3D model for 20 epochs to numerically predict the EF of an echocardiogram, using a mean squared error loss, selecting the model with the highest validation R^2 score. We then test the model on the test set. Comparing our predicted values of EF to the actual EF values, we obtain an R^2 score of 0.55, a MAE of 5.80, and an RMSE of 7.76 between the two quantities.

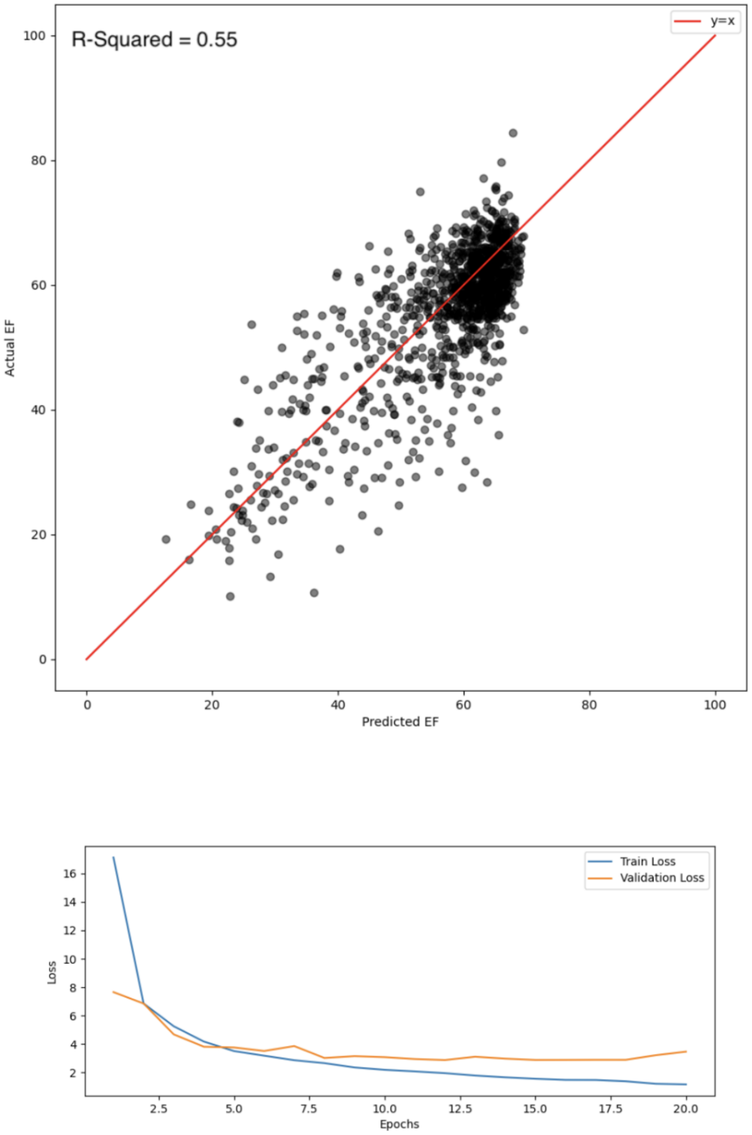


Figure 8: Results of Swin Numberical Prediction Task

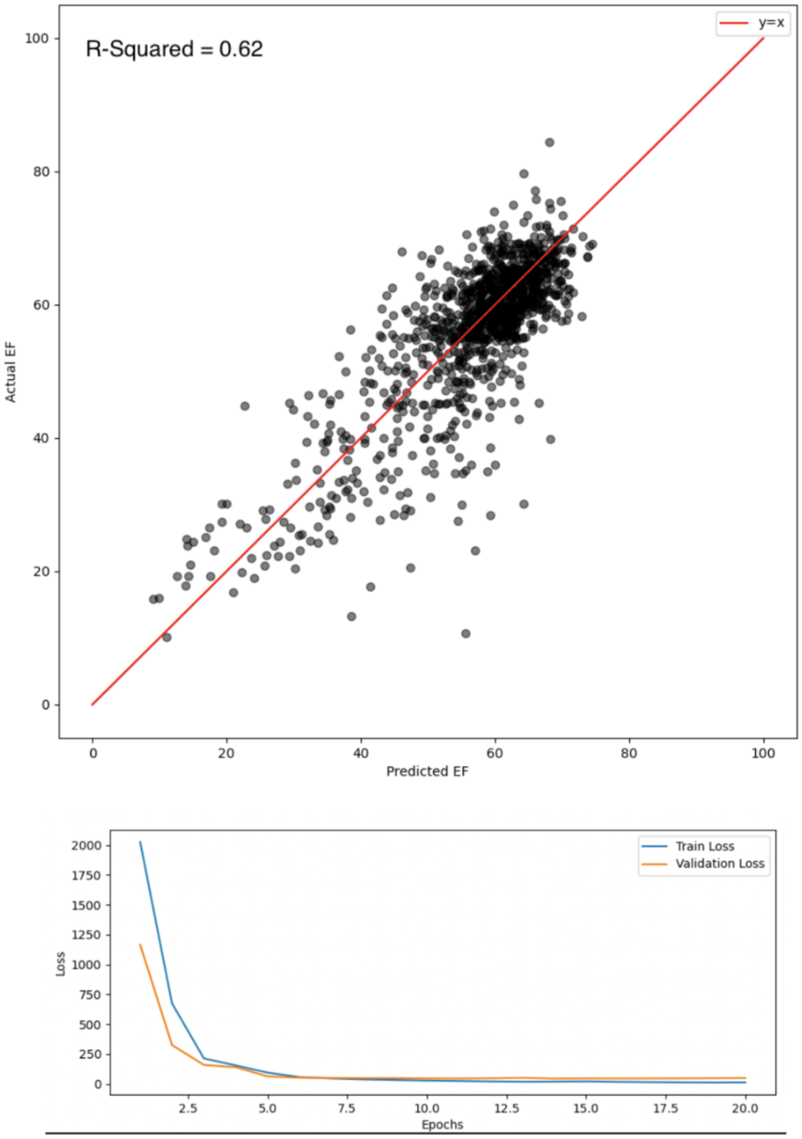
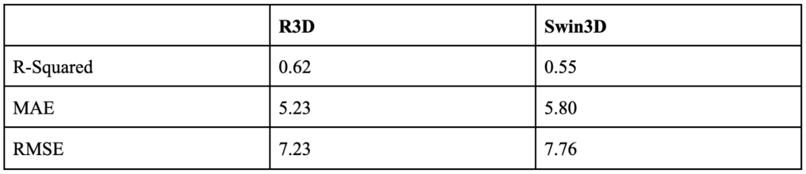


Figure 9: Results of R3D Prediction Task

5.4. R3D for Numerical Prediction

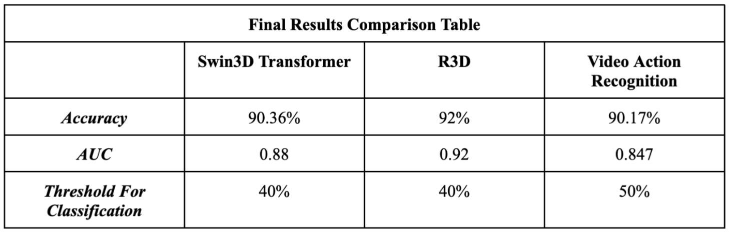
We train the R3D model for 20 epochs to numerically predict the EF of an echocardiogram, using a mean squared error loss, selecting the model with the highest validation R^2 score. We then test the model on the test set. Comparing our predicted values of EF to the actual EF values, we obtain an R^2 score of 0.62, a MAE of 5.23, and an RMSE of 7.23 between the two quantities.

Table 3: Comparison of Swin and R3D Performance



	R3D	Swin3D
R-Squared	0.62	0.55
MAE	5.23	5.80
RMSE	7.23	7.76

Table 4: Overall Comparison of Models



Final Results Comparison Table			
	Swin3D Transformer	**R3D**	**Video Action Recognition**
Accuracy	90.36%	92%	90.17%
AUC	0.88	0.92	0.847
Threshold For Classification	40%	40%	50%

6. Discussion

In general, the performance of our models is on par to most state-of-the-art model performance without data augmentation. The final best models were trained with a downsampled training set with downsampled videos (1 frame in every 4 frames) due to limited computational resources and time, which will weaken the performance. We believe the models of choice have the potential to achieve higher performance due to their key features (i.e. self-attention, spatiotemporal kernel) , but we were not able to demonstrate it due to time and computational resource limitations. In the field of machine learning, more data generally means better performance, and we therefore expect better performance once we train the model on rebalanced dataset with adding augmented data instead of downsampled data.

6.1. Compare the AUC and Accuracy of the Swin3D Transformer to that in Deep Video Action Recognition (Almadani 2022)

In 2022, Almadani et al. reported their utilization of video action recognition (VAR) neural networks to perform binary classification of echocardiograms [6]. Comparison of the AUC and accuracy of the Swin3D Transformer and R3D model described herein to that of the VAR model used by Almadani et al. is as follows:

The accuracies in the Swin3D Transformer and Deep Video Action Recognition paper are similar, demonstrating high overall classification performance. Our model has an accuracy of 89.05% and AUC of 0.87 using the Swin3D Transformer. "Deep Video Action Recognition Models for Assessing Cardiac Function from Echocardiograms" achieved its highest accuracy of 90.17% with 32 frames on the GSM with BNInception. Its highest AUC of 0.847 was achieved with 16 frames on the GSM with InceptionV3. We also trained on the R3D model which achieved a higher accuracy and AUC of 92% and 0.92 respectively; this outperforms the other two methods, demonstrating an even better overall classification performance and improved ability to distinguish the model's instances into the right classes.

6.2. Model Explainability

Gradient weighted class activation mapping (Grad-CAM) is a method of finding which regions of an image play the biggest role in determining the prediction made by a neural network. Grad-CAM allows visualization of CNNs by generating a heatmap, produced by analyzing each feature map using gradients. In order to understand more about our model and its decision, GradCAM is an appropriate tool. Due to the time limit, we were not able to implement a functional GradCAM function but we plan to as part of future experiments.

6.3. Clinical Implications

The results of this study suggest potential application in addressing real world clinical prediction tasks for heart failure. The sucessful prediction of low EF by deep learning models including the transformer based approach, used in this study, implies the utility of deep learning techniques along with their potential as valuable tools in cardiac care. The higher accuracy for predicting low EF by the R3D and Swin Transformer models, is of paramount importance in managing heart failure. By providing reliable predictions of EF, these models offer a promising avenue for enhancing shared decision-making which is a cruical element in patient centric care. However, clinical decision making in a heart failure patient is a complicated process and usually requires contextualization of EF. For example, implantable cardioverter-defibrillators (ICDs) may be considered in patients with clinical heart failure with an EF ¡=35% for primary prevention of sudden death. Hence accurate and timely identification of patients who might be at increased risk of ICD insertion is detrimental for patient important outcomes. While Swin3D transformer and R3D models demonstrated higher performance in predicting EF ¡ 40% compared to other counterparts, their performance may plausibly be generalizable to predict EF ¡= 35% which could potentially facilitate the process treatment selection in these patient. Likewise, the core architecture used in this study can be extrapolated to predict improvement of left ventricular EF following initiation of medical therapy to decide more rapidly on primary device placement in those unlikely to show an EF increase to ¿=35%. However, it is important to note that while deep learning models may provide valuable insights and predictions, they should be viewed as decision support tools rather than standalone diagnostic tools. Contextualization of EF predictions by a cardiologist is crucial for comprehensive clinical decision-making in a heart failure patient. For example, an EF prediction alone may not capture the entire clinical picture.

Additional factors such as the patient's symptoms, prior response to medical therapy, and individualized treatment goals must be considered in real world clinical practice. However, with integration of deep learning models with the clinical expertise of cardiologists, a more comprehensive and patient-centered approach can be achieved. Nevertheless, this collaborative approach would enable the accurate interpretation of model predictions in light of the patient's overall clinical condition, promoting shared decision-making and individualized treatment strategies. By considering the predictions within the broader context of the patient's unique circumstances, healthcare providers can optimize patient care. Therefore these models have the potential to guide a more targeted approach to treatment selection addressing care gaps for high-risk individuals, and potentially reducing the need for costly and unnecessary procedures.

7. Conclusion

The interpretation of imaging is a key component in the diagnosis of heart failure, in particular the use of echocardiography. In this paper, we adapt, train, and deploy R3D and Swin Transformer deep video action recognition networks for the objective of classifying whether or not a heart is healthy based on its ejection fraction calculated using echocardiographic video. We achieve high AUC and Accuracy scores with both these models, and a Mean Absolute Prediction Error of 5.23 and 5.80 respectively on prediction Ejection Fraction from 0 to 100. Future works will likely include using Grad-CAM and other interpretability techniques to analyze our models, data augmentation techniques to increase our training set size, and the use of semantic segmentation models to segment the right ventricle and use this as an input feature to calculating ejection fraction.

8. Acknowledgments and Member Contributions

We are very grateful to our mentors, Prabhu Sasankhan, Po-Chih Kuo, George Tang, and Jacques Kponopu, for their continued guidance and support on clinical relevance, problem selection, model development and training methodologies throughout the project.

Member Contributions: V. Kaker: Wrote majority of ML pipeline code (data preprocessing, model training, model evaluation), proposed/implemented model architectures, applied and obtained access to Supercloud, ran experiments, wrote Results and Methodologies; D. Chung: Worked closely with V. Kaker on ML pipeline code, writing data preprocessing code and proposing/implementing model architectures, wrote Cohort and Methodologies; Y. Zhao: Wrote data augmentation code, worked on Grad CAMs, wrote Discussion section; M. Lee: Worked on Attention Masks to batch inputs, worked on Grad CAMs, wrote Discussion section; S. Perera: Worked on clinical aspects/relevance of project with mentors, organized report writing, wrote Introduction + Related Works; I. Riaz: Worked on clinical aspects/relevance of project with mentors, wrote Introduction + Related Works

References

1. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. Circulation. 2020;141(9):e139-596.

2. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation. 2019;139(10):e56–528.

3. A. Almadani, A. Shivdeo, E. Agu and J. Kpodonu, "Deep Video Action Recognition Models for Assessing Cardiac Function from Echocardiograms," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 5189-5199, doi: 10.1109/BigData55660.2022.10020947.

4. Hunt SA, Abraham WT, Chin MH. 2009 Focused Update Incorporated Into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the International Society for Heart and Lung Transplantation. Circulation 2009;119:e391–479

5. Ouyang, David, Bryan He, Amirata Ghorbani, Matthew P. Lungren, Euan A. Ashley, David H. Liang and James Y. Zou. "EchoNet-Dynamic: a Large New Cardiac Motion Video Data Resource for Medical Machine Learning." (2019).

6. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, Heidenreich PA, Harrington RA, Liang DH, Ashley EA, Zou JY. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020 Apr;580(7802):252-256. doi: 10.1038/s41586-020-2145-8. Epub 2020 Mar 25. PMID: 32269341; PMCID: PMC8979576.

7. Asch FM, Poilvert N, Abraham T, Jankowski M, Cleve J, Adams M, Romano N, Hong H, Mor-Avi V, Martin RP, Lang RM. Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert. Circ Cardiovasc Imaging. 2019 Sep;12(9):e009303. doi: 10.1161/CIRCIMAGING.119.009303. Epub 2019 Sep 16. PMID: 31522550; PMCID: PMC7099856.

8. Thomas A Foley, Sunil V Mankad, Nandan S Anavekar, Crystal R Bonnichsen, Michael F Morris, Todd D Miller, and Philip A Araoz. Measuring left ventricular ejection fraction - techniques and potential pitfalls, 2012.

9. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation. 2018 Oct 16;138(16):1623-1635. doi: 10.1161/CIRCULATION-AHA.118.034338. PMID: 30354459; PMCID: PMC6200386.

10. Vafaeezadeh, M, Behnam, H, Hosseinsabet, A, Gifani, P. CarpNet: Transformer for mitral valve disease classification in echocardiographic videos. Int J Imaging Syst Technol. 2023; 1- 10. doi:10.1002/ima.22885

11. Dai W, Li X, Ding X, Cheng K-T. Cyclical Self-Supervision for Semi-Supervised Ejection Fraction Prediction from Echocardiogram Videos. 2022.

12. Hwang, IC., Choi, D., Choi, YJ. et al. Differential diagnosis of common etiologies of left ventricular hypertrophy using a hybrid CNN-LSTM model. Sci Rep 12, 20998 (2022). https://doi.org/10.1038/s41598-022-25467-w

13. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. and Hu, H., 2022. Video swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3202-3211).

14. Actional Classification on Kinetics 400. Accessible at: https://paperswithcode.com/sota/action-classification-on-kinetics-400

15. Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 3154– 3160, 2017.

16. Tran, Du and Wang, Heng and Torresani, Lorenzo and Ray, Jamie and LeCun, Yann and Paluri, Manohar. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. 6450-6459. 10.1109/CVPR.2018.00675

17. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K. (2018). Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision, ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11219. Springer, Cham. https://doi.org/10.1007/978-3-030-01267-0_19.

18. Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0