


Summer 7-8-2019

Large Scale Electronic Health Record Data and Echocardiography Video Analysis for Mortality Risk Prediction

Alvaro Emilio Ulloa Cerna
University of New Mexico

Follow this and additional works at: https://digitalrepository.unm.edu/ece_etds

 Part of the [Biomedical Engineering and Bioengineering Commons](#), [Computer Engineering Commons](#), and the [Electrical and Computer Engineering Commons](#)

Recommended Citation

Ulloa Cerna, Alvaro Emilio. "Large Scale Electronic Health Record Data and Echocardiography Video Analysis for Mortality Risk Prediction." (2019). https://digitalrepository.unm.edu/ece_etds/467

This Dissertation is brought to you for free and open access by the Engineering ETDs at UNM Digital Repository. It has been accepted for inclusion in Electrical and Computer Engineering ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact amywinter@unm.edu.

Alvaro Emilio Ulloa Cerna

Candidate

Electrical and Computer Engineering

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Marios Pattichis

, Chairperson

Constantinos Pattichis

Manuel Martinez-Ramon

Brandon Fornwalt

Large Scale Electronic Health Record Data and Echocardiography Video Analysis for Mortality Risk Prediction

by

Alvaro Emilio Ulloa Cerna

Electrical Engineer, Pontificia Universidad Católica del Perú, 2010

M.S., Electrical Engineering, University of New Mexico, 2013

M.S., Statistics, University of New Mexico, 2016

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Engineering

The University of New Mexico

Albuquerque, New Mexico

July, 2019

Dedication

to my lovely wife, Jessica.

Acknowledgments

To Prof. Pattichis for his continuous support and guidance.

To Prof. Fornwalt and Prof. Haggerty for their advisement and contributions.

To the DISI lab for the productive discussions.

To Geisinger for the institutional support to research and innovation.

Large Scale Electronic Health Record Data and Echocardiography Video Analysis for Mortality Risk Prediction

by

Alvaro Emilio Ulloa Cerna

Electrical Engineer, Pontificia Universidad Católica del Perú, 2010

M.S., Electrical Engineering, University of New Mexico, 2013

M.S., Statistics, University of New Mexico, 2016

PhD, Engineering, University of New Mexico, 2019

Abstract

Electronic health records contain the clinical history of patients. The enormous potential for discovery in such a rich dataset is hampered by their complexity. We hypothesize that machine learning models trained on EHR data can predict future clinical events significantly better than current models. We analyze an EHR database of 594,862 Echocardiography studies from 272,280 unique patients with both unsupervised and supervised machine learning techniques.

In the unsupervised approach, we first develop a simulation framework to evaluate a family of different clustering pipelines. We apply the optimized approach to 41,645 patients with heart failure without providing any survival information to the underlying clustering approach. The model separates patients with significantly

different survival characteristics. For example, in a 10-cluster model, the minimum and maximum risk clusters had a median survival of 22 and 53 months respectively.

In the supervised approach, with 723,754 videos available from 27,028 unique patients, we assess the predictive capacity of Echocardiography video data for one-year mortality. Also, we hold out a balanced dataset of 600 patients to compare the model performance against cardiologists. We found that the best model, among four candidate architectures, is a 3D dyadic CNN model with an average AUC of 0.78 for a single parasternal long axis view. The model yields an accuracy of 75% (AUC of 0.8) on the held-out dataset while the cardiologists achieve 56% and 61%. The model performance was significantly higher than that of the cardiologists ($p = 4.2 \times 10^{-11}$ and $p = 6.9 \times 10^{-7}$).

Finally, we develop a multi-modal supervised approach that enables interpretability. The model provides interpretations through polynomial transformations that describe the individual feature contribution and weights the transformed features to determine their importance. We validate our proposed approach using 31,278 videos from 26,793 patients. We test our proposed approach against logistic regression and non-linear and non-interpretable models based on Random Forests and XGBoost. Our results show that the proposed neural network architecture always outperforms logistic regression models while its performance approximates the other non-linear models. Overall, our multi-modal classifier based on 3D dyadic CNN and the interpretable neural network outperforms all other classifiers (AUC=0.83).

Contents

List of Figures	xi
List of Tables	xvi
1 Introduction	1
1.1 Electronic Health Records	1
1.2 Interpretability and Explainability	3
1.3 Thesis Statement	4
1.4 Contributions	5
1.5 Organization	5
2 Unsupervised EHR Clustering	7
2.1 EHR Data Simulation	9
2.2 EHR Clustering Pipeline	9
2.3 Simulation Setup	11
2.4 Dataset	13

Contents

2.4.1	Processing pipeline	14
2.5	Statistical analysis	17
2.6	Optimal Pipeline Simulation Results	17
2.6.1	Robustness Experiments	18
2.6.2	Interaction Experiments	19
2.7	EHR Application Results	20
2.8	Discussion	22
2.8.1	Pipeline Optimization	22
2.8.2	EHR Application	25
2.9	Conclusion	27
3	Echocardiography Video Processing	30
3.1	Image Collection and Preprocessing	33
3.2	Electronic health record data preprocessing	34
3.3	Data pruning	35
3.4	Model selection	37
3.4.1	Effect of adding optical flow inputs	39
3.5	Training Procedure	41
3.6	Cardiologist Survey Dataset	42
3.6.1	Software for cardiologist survey	43
3.6.2	Statistical analysis: Machine vs Cardiologists	44

Contents

3.7	Results and Discussion	44
3.8	Conclusion	48
3.9	Future work	50
4	Multimodal Interpretable Risk Prediction	55
4.1	Low Number of Parameters Networks	57
4.2	Interpretability and Explainability	59
4.3	Experimental Setup	62
4.4	Interpretable Neural Network	62
4.4.1	Feature Importance	64
4.4.2	Direction of effect	66
4.4.3	Multimodal Assessment	66
4.4.4	Training, Validation, and Testing	68
4.5	Dataset	69
4.5.1	Electronic Health Records	69
4.5.2	Echocardiography videos	71
4.5.3	Clinical and Video Data Merge	71
4.6	Results and Discussion	75
4.6.1	Significant features	75
4.6.2	Risk model assessment for individual features	76

Contents

4.6.3	A fully interpretable Neural Network based on the top-5 clinical data features	79
4.6.4	Model results	80
4.7	Conclusion	82
5	Conclusion and Future Work	86
5.1	Limitations and Future Research	87
5.2	Conclusion	88
	Appendices	90
	A Supplementary Tables	91
	References	101

List of Figures

2.1	Missingness experiment results.	18
2.2	Robustness experiments results for (a) Effect size, d , (b) number of informative features m , and number of noisy features η	20
2.3	Kaplan Meier curves for the Highest and Lowest risk clusters compared to HFrEF and HFpEF for (a) 2, (b) 10, (c) 30, and (d) 50 clusters.	21
2.4	t-SNE visualization of 60 encoding units from Autoencoder. In each row, the left two plots are 2D-density plots for each category in the right plot, where each dot is a patient with Heart Failure. The same visualization is colored by (a) the Aortic Root diameter, (b) Left Ventricular Ejection Fraction, (c) Hypertension, and (d) Ischemic heart disease.	28
2.5	Error distribution for Autoencoder with 2 layers (500 hidden units each) and 60 encoding units	29

List of Figures

2.6	Interaction experiments for (a) Effect size vs Noise, and (b) Effect size vs Missingness. The gray areas denote where neither method scored above 0.8, the white areas denote no significant difference between score means. The colored areas denote significant differences between LLE and DAE.	29
3.1	Examples of raw (left) and annotated (right) videos.	33
3.2	Number of patients for experiments that required 3, 6, 9, and 12 months follow-up (as indicated in the Extended Data Table 2) with the proportion of dead patients (shaded bar).	36
3.3	Neural network architecture for mortality prediction from echocardiography videos and electronic health record (EHR) data. The convolutional layer (Conv) is shown on the top box with a solid outline and the tabular layer (Tab) is shown in the bottom box with a dashed outline. The convolutional layer consists of Convolutional Neural Networks (CNN), Batch Normalizations (Batch Norm.), rectified linear units (ReLU), and a three-dimensional Maximum Pooling layer (3D Max Pool). The tabular layer consists of a fully connected layer (Dense) with sigmoid activations and a Drop Out layer. The input video dimensions were 150 x 109 x 60 pixels, and the output dimension of every layer are shown.	39
3.4	AUCs of one-year mortality predictions across all views with four different neural network architectures: 2D CNN + Global Average Pooling (GAP; dark gray), 2D CNN + Long Short-Term Memory (LSTM; light gray), a 3D CNN + GAP (light blue), and 3D CNN (dark blue).	40

List of Figures

3.5	AUCs of one-year mortality predictions across all views with different levels of reduced resolution ranging from native (x1) to 4-fold (x4). Note that full native resolution training was only done for select views due to the computational time required to complete the experiment at this resolution.	40
3.6	One-year mortality prediction performance ranking for all echocardiography views using only the raw video (blue) versus the raw video with optical flow features (gray).	41
3.7	Interface of the web application developed for cardiologists to predict survival one year after echocardiography	43
3.8	Mortality prediction performance for echocardiographic videos alone at 3, 6, 9 and 12 months for all views. The error bars denote one standard deviation above and below the average across 5 folds. . . .	46
3.9	Cardiologists vs Machine performance for 1-year mortality prediction from the survey dataset of 600 samples with balanced prevalence. The left plot (a) shows the accuracy in bars and sensitivity (red) and specificity (green) as triangles. The right plot (b) shows the operating points of the cardiologists as orange dots, the Receiver Operating Characteristic curve for the machine performance in blue, and the machine operating point as a blue dot.	47
3.10	Learning curves for the full (158) EHR variables model compared to the full EHR variables plus videos. The AUC is reported on the 600 patient set as a function of training set size, ranging from 10 to the maximum number of datasets available for the given data inputs, which was 501,449 for the EHR variables and 26,428 for the Full EHR+videos.	49

List of Figures

- 4.1 General framework for multi-modal risk assessment. EHR data and cardiac ultrasound videos are input to the risk assessment system. We emphasize the use of separable non-linear models where we look at contributions from each modality and each feature separately and also within the joint multi-modal framework. The mortality risk assessment is used to inform treatment. 56

- 4.2 Data flow from input to risk score calculation for the proposed multi-modal system. The input is based on an Echocardiography exam and other clinical information (height, weight, etc.). The physician/technician then reads and generates measurements from the video and clinical data from the patient’s exam. The output of the 3D CNN video analysis system is connected directly to the final layer of the model. The measurements and clinical data are transformed with the proposed Interpretable Neural Network (INN) which learns 3rd order polynomial transformations that can then contribute to the final risk score. ConvNet[1,2,3] are described in Table 3.4. 73

- 4.3 Example of low-level features extracted from the parasternal long axis view. In this example, we show (a) a frame of the input video and (b) all outputs from the four feature maps produced by L1 (top), L2 (middle up), L3 (middle down), and the flatten layer enhanced for visualization (spans 10 rows). 74

List of Figures

4.4	Full model risk functions (blue) with normalized histograms of survivors (light green) and non-survivors (red orange). When the two histograms overlap, the histograms appear light brown. Risk function for (a) Age in years, (b) heart rate in beats per minute, (c) weight in kilograms, (d) diastolic and (e) systolic blood pressure in mm Hg, (f) left ventricular ejection fraction in percent, (g) Tricuspid regurgitation maximum velocity in cm/s., (h) aortic insufficiency deceleration slope (AI dec slope) in cm/s ² , (i) left ventricular internal dimension at end-diastole in cm, and (j) left ventricular end systolic volume in ml. The uncertainty in the risk functions are derived from the 5 results across the 5 runs.	84
4.5	AUC performance as a function of the number of the most significant input features for clinical data (left) and echocardiography video measurements (right) for Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), and the proposed method (INN).	85

List of Tables

2.1	Simulation experiments with default parameters $p = 0$, $\eta = 0$, $d = 10$, $n = 5000$, and $m = 10$	11
2.2	Baseline results for identifying best scaling for each feature reduction method . The entries show average score and standard deviation of the scores across repetitions.	13
2.3	Summary of the historical Echocardiography records at Geisinger and affiliations. Gender was not reported in 172 studies, and the status of the patient in 761 studies is unknown. Age is shown as mean±standard deviation. The diagnosis of Heart Failure follows the eMERGE guidelines [1]	14
3.1	AUC scores for each data modality combination of EHR and Echo video data on the 600 left out studies used to compare to the cardiologists. No video models were trained on all available studies, whereas Single Video and All Videos were trained on a subset where video data were available. The No EHR variables and No Video cell denotes a random guess.	48
3.2	Low-parameter 2D CNN + LSTM with 4,237 trainable parameters .	51

List of Tables

3.3	Low-parameter 2D CNN + GAP with 3,477 trainable parameters . . .	52
3.4	Low-parameter Dyadic 3D CNN with 14,309 trainable parameters. For the dimension of the 3D CNN architecture, refer to Fig. 4.2. . . .	53
3.5	Low-parameter Dyadic 3D CNN + GAP with 13,685 trainable pa- rameters. For the dimension of the 3D CNN architecture, refer to Fig. 4.2.	54
4.1	Comparison of Parameters/Cases ratio of different ImageNet models: (i) AlexNet [2], (ii) Inception V3 [3], (iii) DenseNet [4].	58
4.2	Demographics table of 31,278 EHR samples.	69
4.3	Top five features with their corresponding coefficients for different models. For each coefficient, we provide 95% confidence intervals in between square brackets. Confidence intervals were computed using five folds.	74
4.4	Normalization parameters for eq. (4.13). The minimum and max- imum values are used for normalizing the input so that it varies between -1 and +1. The transformation equation is described in eq. (4.12).	78
4.5	Models performances in percent AUC units. For each method and data combination, we present the average AUC and standard devi- ation based on 5 independent runs. We use the term Interpretable Neural Network (INN) to refer to the proposed method. The CD input does not include EVM or Video features.	82
4.6	Performance of 2D and 3D CNN video models in percent units for single and multi-modality inputs.	82

List of Tables

A.2	Description of all variables extracted from the electronic health records. ■ *MOD = modified ellipsoid, **el = (single plane) ellipsoid, LV = left ventricular, IV = inter-ventricular. ¹⁻¹⁰ Selected EHR variables pre- viously reported as the top 10 predictors of 1-year mortality. *Hot encoded for severity levels 0,1,2,3. Diastolic function coding -1: Nor- mal, 0: abnormal (no grade reported), [1,2,3]: grade I/II/I	91
A.3	Number of valid samples after setting 600 studies aside for the final test comparison to the 2 cardiologists.	98
A.1	View labels found in DICOM tags for the corresponding view type. The view tag in bold indicates the abbreviation used for the view type.	100

Chapter 1

Introduction

1.1 Electronic Health Records

Recent advances in technology and medicine have enabled the implementation of Electronic Health Records (EHR) to facilitate patient management. The EHR data contain a patient's full clinical history including imaging measurements, demographics, laboratory values, and treatments.

Comprehensive interpretation of this rich information should support physicians to optimize predictions about the most appropriate diagnosis, prognosis and treatments. However, the complex and heterogeneous nature of this rich dataset preclude physicians from fully digesting all the information [5]. Thus, in the current clinical practice, diagnoses are only based on a few pieces of information, and are often times broad and generic [6].

Risk calculators, based on metrics from the patient's health, are widely spread in clinical use. For example, the Framingham risk score [7] yields the risk of developing a cardiovascular disease within ten years, and the Seattle Heart Failure score [8]

Chapter 1. Introduction

predicts 1-, 2-, and 3-year mortality in patients with Heart Failure. Yet, these scores are mainly developed in a controlled environment which is not a realistic scenario for a real world application. In consequence, these current risk scores yield poor generalization [9, 10, 11].

Leveraging the complex data in EHR is non-trivial. Different components of patient data (e.g. imaging data, demographics, laboratory values, etc.) often live in different tables of the databases; human errors are often involved during data entry, which results in unrealistic measurements as human errors are often involved during data entry. For example, different units may be used for the same measurements; multiple clinical tests might be ordered to obtain the same information, or the same test may be performed multiple times (redundant information); many fields in the imaging report may be sparsely filled depending on the purpose of the study.

Fortunately, advances in data analysis and machine learning can be harnessed to train computers to fully utilize the information from EHR and make a more informative and personalized prediction of a patient's risk. We hypothesize that risk models trained on EHR data can use the historical EHR information to produce an accurate prediction of future clinical events superior to state-of-the-art clinical risk models.

In particular, deep learning has gained popularity because of the performance yields in multiple fields of study [12]. These complex neural network structures are most effective on large datasets and has shown promise in natural images, video, audio, and text. The advances in neural networks amount from the development of regularization techniques [13], gradient descent algorithms [14, 15], and the reduction of parameters [2].

In the medical field, deep learning models dominate research on medical images [16] and tabular EHR data [17, 18]. To particularly highlight the capabilities

Chapter 1. Introduction

of deep learning in medical imaging analysis, such methods have been used in ultrasound video analysis for frame labeling tasks such as segmentation of certain chambers of the heart (the left ventricle) [19, 20], and fetal standard image plane / orientation detection [21, 22].

The rapid rise of deep learning methods has also been associated with the development of lower-parameter neural network systems that can also deliver better performance than previously considered methods. To demonstrate the trend, we consider some of the most popular successful classifiers. In 2012, AlexNet used 60M parameters to achieve a top-5 test error rate of 15.3% for the ILSVRC-2012 competition [2]. In 2016, the updated version of the Inception architecture used about 25M parameters to achieve a top-5 test error rate of 5.6% for the same competition [3]. In 2017, DenseNet-201 used 20M parameters model to achieve 6.34% accuracy on the same dataset and thus match the performance of a 101-layer ResNet with more than 40M parameters [4].

1.2 Interpretability and Explainability

Linear models such as linear or logistic regression, are inherently interpretable models. When the input variables are equally scaled, the coefficients of the linear predictor can be used to assess feature importance based on the coefficient's magnitude, and the effect directionality based on the coefficient's sign. Some examples of clinical adoption of linear models are the Framingham risk score [23], and the Seattle Heart Failure score [8]. Unfortunately, the performance of linear models can be limited [24].

A well calibrated non-linear model can outperform any linear model at the cost of interpretation ability. While not as direct as linear models, there are also approaches to support explanations for non-linear methods. As described in section 4.2, explanation models build around black-box models and approximate simpler in-

Chapter 1. Introduction

interpretable models at the cost of performance, whereas interpretable models do not require separate explanation models.

Explanations are a step forward to interpretability, however inherently interpretable models are still desired in high stake decisions such as the prediction of clinical outcomes [25]. An interpretable model would allow understanding of the data source, which at the same time enables us to detect any irregularities and iterate on the input data processing. Moreover, the ability to detect issues in the data source would only help improve performance.

When a non-interpretable model latches on confounding factors, such as artificial annotations in an image [26], it would possibly escape from the cross-validated performance metrics. However, an interpretable model would have shown that the confounding factor play an important role in the final decision. Furthermore, explanation models forces us to relay on two models (the black-box and explanation model), which, by design, disagree with each other. If they were to agree all the time then the explanation model would be preferred.

In this dissertation, I designed inherently interpretable neural network which once trained, yields inter-modality feature importance, feature response functions, and intuitive interpretations.

1.3 Thesis Statement

Given the potential of EHR data to inform and describe the patient’s health status, I hypothesize that incorporating multiple EHR sources to calculate the patient risk, with interpretability while maintaining the performance of best performing models, could yield accurate and transparent models for clinical use.

1.4 Contributions

The primary contributions of the dissertation include:

- *Suitable EHR clustering pipeline:* I propose a simulation framework and use it to evaluate multiple clustering pipelines. We apply the pipeline to patients with Heart Failure, and ultimately find a characterization that separates risk better than the clinical standard of Heart Failure with preserved or reduced Ejection Fraction.
- *Optimal deep learning architecture for mortality prediction from Echocardiography videos:* I design and report an optimal deep learning architectures for Echocardiography video analysis, and present the potential of combining video and tabular data for mortality prediction.
- *Multi-modal Interpretable Neural Network architecture:* I propose an inherently interpretable neural network that has the ability to rank multiple data modalities while learning transformation function for individual tabular inputs. The rank and the transformation functions compose a transparent model that does not sacrifice performance when compared to other non-linear, non-interpretable models.

1.5 Organization

The remainder of the dissertation is organized as follows:

- In chapter 2, I describe the EHR simulation framework used to obtain a suitable clustering pipeline, and its application to patients with Heart Failure.

Chapter 1. Introduction

- In chapter 3, I describe the experimental design for the search of an optimal neural network architecture for Echocardiography video classification.
- In chapter 4, I describe the proposed inherently interpretable neural network and its application to patients with available Echocardiography video data.
- In chapter 5, I summarize the research findings, state concluding remarks and future work.

Chapter 2

Unsupervised EHR Clustering

Doctors provide diagnoses to help predict the health trajectory of their patients. A diagnosis also helps to predict what treatments have the highest likelihood of improving a patient's health. The more granular the diagnosis, the more specific or “precise” medicine can become. The wealth of medical data gathered from patients, that is digitally available in an electronic health record (EHR), should support highly granular diagnoses. Unfortunately, the current clinical paradigm of a human physician wading through this vast sea of data cannot deliver the promise of precision medicine.

Fortunately, advances in machine learning can be harnessed to sift through this rich dataset and extract useful information to facilitate human decisions. One popular application is phenotyping by cluster analysis. Previous studies [27, 28, 29] have shown that clustering algorithms have the potential to classify patients into similar phenotypes based on data contained in the medical record. For example, using unbiased hierarchical cluster analysis and penalized model-based clustering, Shah et al. [28] identified 3 phenotypes in patients diagnosed with heart failure with preserved ejection fraction. Upon identification of such granular and more homoge-

Chapter 2. Unsupervised EHR Clustering

neous clusters, the outcomes (e.g. hospitalization, cardiac events or mortality) and attempted therapies within each cluster can then be linked together to predict likely outcomes resulting from choosing particular therapies.

While there have been many advances in the field of cluster analysis [30], the methods rely on the assumption of homogeneous, non-redundant and complete data. However, EHR data are heterogeneous (variables can be continuous or categorical, and with different scales), redundant (multiple measurements may assess the same underlying patient feature), incomplete (many fields in the clinical reports are sparsely filled depending on the purpose of the study), and noisy (not all variables are informative in all conditions). Additionally, human errors and system biases also contribute to measurement errors in EHR data. Thus, to fully utilize the EHR to reliably detect disease subtypes, clustering techniques must be paired with pre-processing techniques that normalize and reduce the complexity of the raw EHR data. Such a clustering pipeline, including pre-processing steps, has not been previously proposed or validated.

In this chapter, we assess, propose, and apply the optimized clustering pipeline that is robust to the nuisances of EHR data. The pipeline consists of imputation, normalization, feature reduction, and clustering. Multiple commonly used techniques are evaluated at each step, and the best performing pipeline is selected. Since the accuracy of clusters in real EHR applications cannot be measured due to lack of a ground truth, we assessed accuracy using simulated EHR data where ground truth could be easily defined. To the best of our knowledge, this is the first study to propose and validate an unsupervised homogenization pipeline for EHR clustering.

2.1 EHR Data Simulation

We simulated patient encounters with a sample generator that mimics the redundancy and heterogeneity of EHR data. We defined rows for patient encounters (samples) and columns for measurements taken from the patient (features). We designed three clusters with n samples per cluster, observed dimensionality m , and effective dimensionality of 2 (for visualization convenience).

The sample generator drew $3n$ independent samples from a multivariate normal distribution with $\mu = [0, 0]$, and $\Sigma = \mathbf{I}_{2 \times 2}$ to form the matrix $N_{3n \times 2}$. Then, we separated the clusters by shifting n samples at a time. The first n samples stayed in the origin, while the next n were shifted by $[d, 0]$, and the last n were shifted by $[\frac{d}{2}, \frac{\sqrt{3}}{2}d]$, forming an equilateral triangle with a distance d from each vertex.

We emulated redundancy by projecting the original feature vector to a m dimensional space: $X_{3n \times m} = N_{3n \times 2}P_{2 \times m}$, where the elements of the projection matrix, P , were drawn from a uniform distribution in the range $(0, 1)$.

We then enforced heterogeneity by quantizing half of the variables (set to zero if below the mean and 1 otherwise), chosen at random, and scaling each continuous feature with a random factor between 1 and 100. Finally, we added Gaussian noise ($\mu = 0, \sigma = 1$) to every element in the data matrix to mimic measurement errors.

2.2 EHR Clustering Pipeline

Imputation

We tested median imputation, where the median value from valid samples complete missing values; k-Nearest Neighbors (KNN), where the average value from the k-

Chapter 2. Unsupervised EHR Clustering

nearest samples is used; and Multiple Imputation by Chained Equations (MICE) [31], where the missing values are predicted based on regression models with complete samples.

Normalization

For continuous variables, we tested Z-score, where every variable is set to zero mean and unit variance; MinMax, which normalizes to a [0,1] range; and Whitening, where the feature space is linearly projected such that inter-feature covariance is the identity matrix.

Feature reduction

We propose the use of Deep Autoencoders (DAE) [32] and Denoising Autoencoders (DnAE) [33] for EHR feature reduction. Autoencoders are trained to reconstruct an input through encoding and decoding networks. In the DnAE case, noise is added to the encoded units to enforce robustness to measurement noise.

We designed the network architecture with a hyper-parameter search for layers, hidden, and encoding units. The network with the least number of encoding units that achieves the reconstruction error of 1% or less is preferred. The encoding vectors represent EHR data in a compressed and continuous vector, suitable for any clustering technique.

For comparison, we evaluated other methods with local (Local Linear Embedding (LLE) [34]) and global neighbor algorithms (Isometric Mapping (ISOMAP) [35]); as well as affinity matrix algorithms, such as Spectral Embedding [36] and Multidimensional Scaling (MDS) [37].

Table 2.1: Simulation experiments with default parameters $p = 0$, $\eta = 0$, $d = 10$, $n = 5000$, and $m = 10$.

Experiment	Parameter	Levels
Effect Size	d	[3, 4, 5, 6, 7, 8]
Features	m	[6, 20, 40, 100, 200, 500]
Missingness (%)	p	[0, 10, 20, . . . , 80]
Noise	η	[4, 16, 64, 128, 256]

Clustering

For simplicity, we used K-means to conduct the final cluster analysis.

2.3 Simulation Setup

First, we simulated a baseline scenario where all parameters were set to an ideal level with complete, free of noise, $d = 10$, 5000 samples per cluster, and $m = 10$. An effect size of 10 resulted in less than 0.01% overlap between clusters, and heuristically $m = 10$ resulted in good performance for all pipelines. This baseline was used to identify the best performing pair of normalization and feature reduction methods, which were then used in the rest of the experiments.

We then simulated four scenarios for testing the pipeline robustness at various levels of severity. In all experiments, we swept one simulation parameter while keeping all others constant. We measured the adjusted rand-score [38], which computes a similarity measure between the results of two sets of labels by counting pairs that are assigned in the same or different clusters in the predicted and true clustering while adjusting for random chance. Table 2.1 describes each experimental setup and the default parameters. Every experiment was run 5 times to extract the mean and standard deviation of the performance.

Chapter 2. Unsupervised EHR Clustering

Missingness

To simulate the missing entries in the EHR, we randomly removed a percentage p , from the observed data matrix and denoted them as missing values. We varied p from 0 to 80% in increments of 10%.

Effect Size

We manipulated the effect size by varying the distance between cluster centers, d . In two dimensions, we can calculate the number of overlapped samples by counting the number of samples beyond a distance of $\frac{d}{2}$ in a bivariate standard normal distribution. Then, in a triangular setting, the number of overlapped samples would be 6 times the calculated amount. By conducting a Monte Carlo simulation, we can convert the effect sizes of 3–8 to the percentage of overlapped samples [13.35%, 4.55%, 1.24%, 0.27%, 0.04%, 0.01%]. This can be interpreted as the lower-bound for error in cluster assignment.

Redundant Features

We assessed the robustness to the number of redundant features present in the dataset by increasing the dimensionality, m , while keeping the ground-truth dimensionality of 2. We simulated projection matrices that generated [6, 20, 40, 100, 200, 500] features.

Uninformative/Noisy Features

EHR data contain information that may not be useful in determining clusters of similar patients. We assessed the effects of including non-informative variables by

Table 2.2: Baseline results for identifying **best scaling for each feature reduction method**. The entries show average score and standard deviation of the scores across repetitions.

	MinMax	Raw	Whitening	Z-score
DAE	0.982(0.03)	0.822(0.20)	0.841(0.15)	0.998(0.00)
DnAE	0.983(0.03)	0.769(0.20)	0.781(0.20)	0.998(0.00)
MDS	0.903(0.17)	0.985(0.03)	0.294(0.18)	0.999(0.00)
ISOMAP	0.264(0.41)	0.235(0.35)	0.822(0.26)	0.390(0.43)
LLE	0.737(0.20)	0.976(0.05)	0.503(0.32)	0.745(0.21)
Spec. Emb.	0.770(0.21)	0.994(0.01)	0.634(0.29)	0.753(0.23)

appending η random continuous and η random binary variables.

2.4 Dataset

In the span of 26 years (1991-2017), Geisinger Health System and affiliations gathered 427,012 Echocardiography studies from 206,650 patients. We extract demographics (see Table 2.3), ICD codes, Echocardiography measurements, clinical notes, and Heart Failure diagnostic [1].

The demographics contain gender, self reported race and smoking status, and date of birth and death, for each patient. The ICD table contain ICD-10 codes with onset and resolved dates. Old ICD-9 codes are mapped to the ICD-10 standard for consistency. The Echocardiography table consists of 528 measurements and session date. The clinical notes are free text notes written by clinicians that describes findings and impression from each session, we only extract the estimation of the left ventricular ejection fraction (LVEF).

Table 2.3: Summary of the historical Echocardiography records at Geisinger and affiliations. Gender was not reported in 172 studies, and the status of the patient in 761 studies is unknown. Age is shown as mean±standard deviation. The diagnosis of Heart Failure follows the eMERGE guidelines [1]

	Gender		Status		Age
	Female	Male	Alive	Deceased	
Not HF	187,049	198,160	289,064	95,542	61.8±17.3
HF	18,412	23,219	21,928	19,717	72.2±12.9
Total	205,461	221,379	310,992	115,259	62.8±17.2

2.4.1 Processing pipeline

This processing pipeline that takes EHR data tables and assigns cluster labels to each Echocardiography study. The pipeline consists of the following steps performed sequentially: merging, cleaning, imputation, feature space reduction (homogenization), and clustering.

Merging

We define a sample as an Echocardiography study, thus all other tables are modified to meet the same format. From the demographics table, we compute the age and include the survival time in months from the study date, including the known status of the patient (deceased or alive). The LVEF values are appended to the table since it has a direct relation to each study.

The reformatted ICD table indicates whether a code is active or not at the time of the study, i.e if the study date falls between the code onset and resolved dates, the code is set to one, zero otherwise. Also, we only include codes relevant to the circulatory system (I codes). Similarly, the HF diagnosis is computed for each Echocardiography study following the eMERGE guidelines [1].

Cleaning

We identify and remove measurement errors by detecting out-of-range values. The pipeline set thresholds based on heuristics and remove any value outside the valid range. For example, LVEF is a measurement that denotes the percentage of blood pumped out of a ventricle of the heart, however the data is prone to typographical errors since clinicians write it in a free text report. Thus, we flag any measurement that is negative or larger than 100 as missing.

None of the measurements in our tables should be negative, we set the minimum to zero. For measurements in which a maximum is not defined, we set the threshold to the average plus three standard deviations as the maximum value.

Imputation

Given the wide array of tests and measurements that can be obtained on patients, missing data is common, as it is unlikely that every patient has every possible test and measurement. Physicians evaluate the cost-benefit of each test and may not request one if normal results or no significant difference from the previous test is expected. Based on this assumption, we can rely on past and future information to interpolate some missing values.

To complete the rest of missing data, i.e when a patient never had a test or measurement, the pipeline uses multiple imputation by chained equations (MICE) [31].

Feature reduction

EHR data is both heterogeneous (i.e. the variables can be continuous, binary, or categorical) and redundant (i.e. multiple tests or measurements may assess nearly the same underlying patient feature). The heterogeneity and redundancy of EHR

Chapter 2. Unsupervised EHR Clustering

data is reduced to a continuous space with deep-autoencoders, where categorical variables are encoded with a one-hot technique, and continuous data is normalized to a 0 to 1 range.

An autoencoder is a tool for extracting latent features without knowledge of pre-conceived labels. Autoencoders train to optimally reconstruct the input through encoding and decoding networks. The encoding network reduces the inputs dimensionality and produces a compressed feature vector through a non-linear mapping. The feature vector is then decoded to reconstruct the original input.

We train a deep-autoencoder for a maximum of 1000 epochs, stopping if there is no reduction in the loss function for 50 epochs. The architecture is designed by conducting a hyper-parameter search. We evaluate the number of layers, number of hidden units, and cost function (mean squared error vs cross-entropy), where the network with the least number of encoding units that achieves the desired reconstruction error is preferred. The encoding units are used to represent EHR data in a compressed and continuous feature vector, suitable for any clustering technique.

Clustering

We use clustering to explore the underlying structure of the encoding units and yield a natural classification of studies. Since the encoding units are continuous and homogeneous, we apply a classical and intuitive clustering technique, K-means [39], which labels k groups of similar patients. The similarity is set as the Euclidean distance between encoding feature vectors.

We use metrics of intra-cluster similarity and inter-cluster separation, such as the silhouette score, to guide the value of k . However, as those metrics have limitations [40] we also evaluate the relationships of the extracted phenotypes to outcomes and optimal therapies for several values of k in order to find the most clinically useful

set of phenotypes.

2.5 Statistical analysis

We used Cox Proportional Hazard Regression (CPH) [41] to predict survival as a function of time. By defining birth as the study date and death as all-cause mortality, we can assess how different phenotypes broadly relate to outcomes.

The new phenotype of patients with HF is compared against the traditional classification, reduced ($\leq 50\%$) or preserved ($> 50\%$) LVEF, computing the cross-validated concordance score from the CPH models that predicts survival from the clusters.

2.6 Optimal Pipeline Simulation Results

The baseline experiment revealed that the performance of the clustering pipeline heavily depended on the choice of normalization and feature reduction method (see Table 2.2). DAE, DnAE, and MDS paired best with Z-scoring, all with scores above 0.99. ISOMAP performed best with Whitening while LLE and Spectral Embedding obtained its best performance when no scaling was used. We used these optimal pairs to conduct the remainder of the experiments.

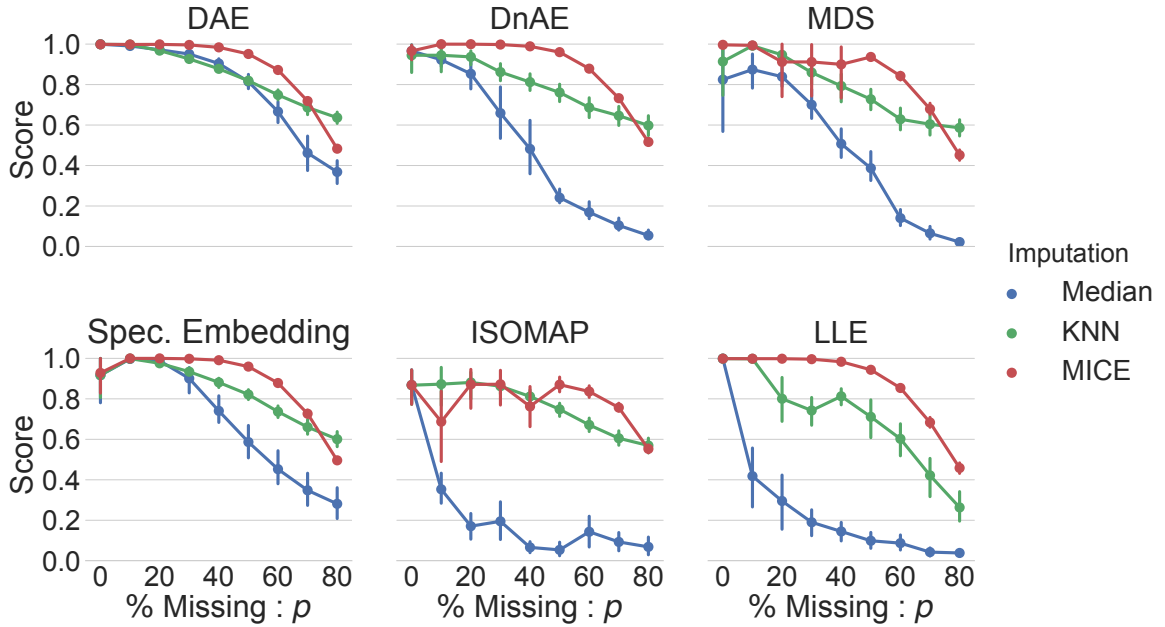


Figure 2.1: Missingness experiment results.

2.6.1 Robustness Experiments

Missingness

As shown in Fig. 2.1, levels of missingness above 60% significantly impaired the clustering performance for all pipeline configurations (all scores below 0.8). Among the three imputation methods, MICE resulted in the best performance for all feature reduction methods except ISOMAP, for which KNN was marginally better up to 50%. Median imputation consistently the worst performance.

Effect Size

As expected, the performance of all configurations increased with the effect size (Fig. 2.2a). Overall, the top three performing feature reduction methods were LLE, DAE, and MDS. LLE exhibited the best performance across feature reduction methods but

Chapter 2. Unsupervised EHR Clustering

only marginally better than DAE, e.g. the p-value of a paired t-test was 0.03 at the effect size of 4.

Features

LLE, DAE, and MDS were essentially immune to large amounts of redundant features (Fig. 2.2b). DnAE appeared to be similarly immune at low levels, but its performance sharply decreased with greater than 200 features. Conversely, Spectral Embedding benefited from higher numbers of redundant features and performed on par to the best methods for 200 and 500 redundant features. ISOMAP performed poorly at all levels

Uninformative/Noisy Features

As shown in Fig. 2.2c, most methods, except DnAE and ISOMAP, were immune to large amounts of uninformative variables. DnAE was robust to uninformative variables up to 32 continuous and binary uninformative variables. ISOMAP did not tolerate even the minimum number of uninformative variables.

2.6.2 Interaction Experiments

Following the robustness experiments, we identified DAE and LLE as the top 2 best performing feature reduction methods overall. To further compare these methods, we performed subsequent experiments that allowed for interactions of varying effect sizes, missingness, and noise.

Overall, LLE matched or outperformed DAE. In the effect size vs noise experiment, (Fig. 2.6a) large amounts of uninformative variables and medium effect sizes favored LLE. In the effect size vs missingness experiment, (Fig. 2.6b) LLE showed

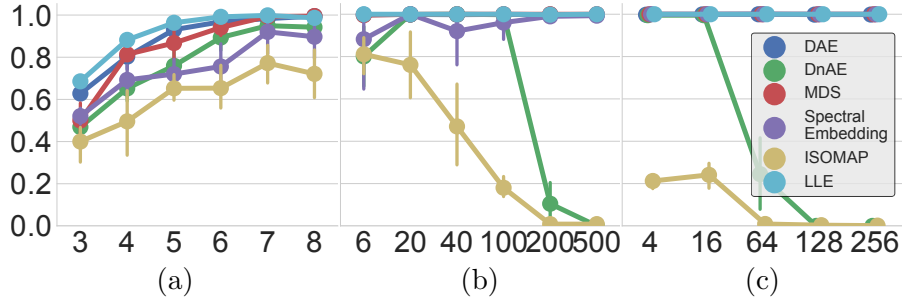


Figure 2.2: Robustness experiments results for (a) Effect size, d , (b) number of informative features m , and number of noisy features η .

significantly better performance for medium effect sizes and low missingness and no difference for large effect sizes and low to medium missingness.

2.7 EHR Application Results

We designed the deep autoencoder following a hyper-parameter search for different network configurations, such as depth (from 1 to 3 hidden layers), number of encoding units (10, 40, 60, and 100), hidden units for each layer (50, 200, and 500), and cost function (cross-entropy and mean squared error). A 3-layer architecture of 200, 200 and 40 hidden units with mean squared error as the cost function resulted in a 1% absolute reconstruction error, which was the minimum encoding size that yielded the desired reconstruction error. Out of all the Echocardiography studies used to train the deep-autoencoder, we extracted the compressed feature vector of 41,647 studies that fit the eMERGE criteria for HF.

We conducted a clustering analysis using K-means on the encoded space. We varied the number of clusters from 2 to 100 and fit a Cox Proportional Hazards Regression model to explain survival time of each patient for each cluster number. The results suggest that the difference in median survival between the highest and

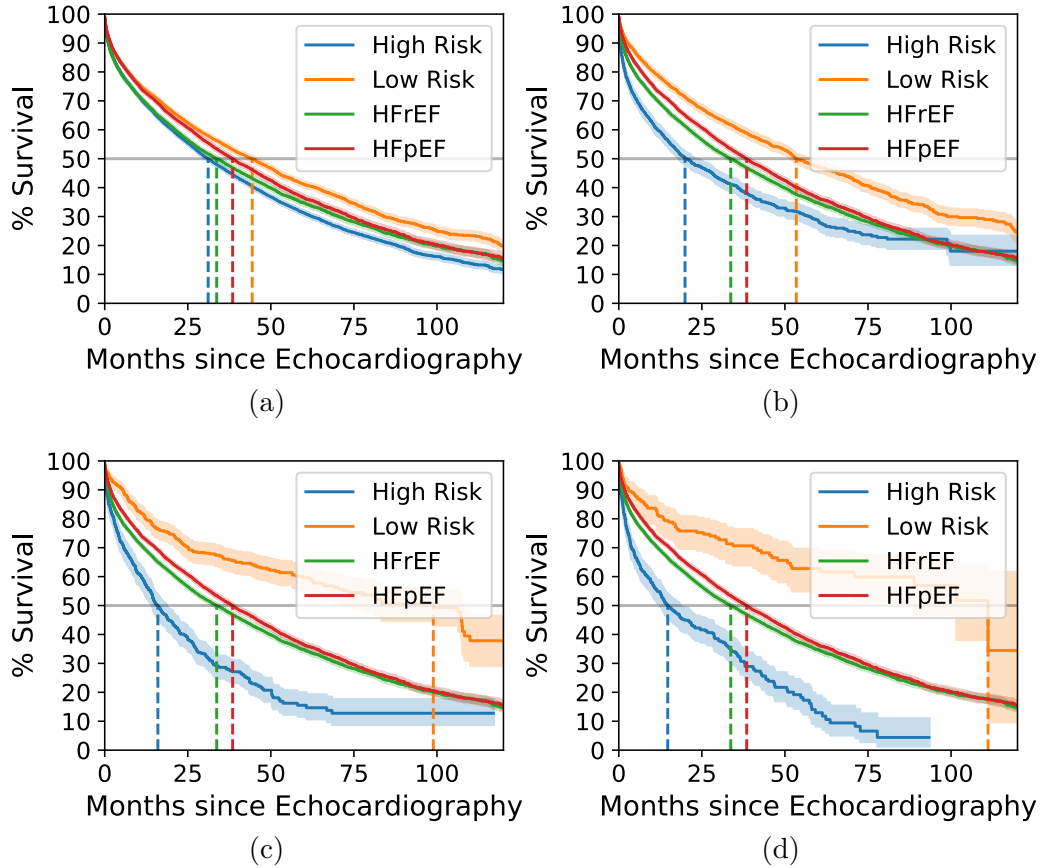


Figure 2.3: Kaplan Meier curves for the Highest and Lowest risk clusters compared to HFrEF and HFpEF for (a) 2, (b) 10, (c) 30, and (d) 50 clusters.

lowest risk groups in the proposed categorization monotonically increased beyond the clinical classification difference that was only 4.5 months (31.8 vs 36.3), see Fig. 2.3.

Given the clinical value of the extracted phenotypes, we visualized the encoded space (40 dimensions) in a 2-dimensional space using t-distributed stochastic neighbors embedding (t-SNE), see Fig. 2.4. We color coded raw features on top of the learned representation and visually assessed what features were most useful for separating the clusters. For example, Fig. 2.4a shows that the aortic root diameter and

ICD code I-25 clearly discriminate the patients into large clusters.

2.8 Discussion

We proposed and evaluated a clustering pipeline tailored for complex EHR data by comparing performances of commonly used techniques. We found two pipelines that outperform other alternatives: 1) MICE imputation + LLE feature reduction; 2) MICE imputation + Z-score normalization + DAE feature reduction. Both pipelines are robust to missingness (up to 60%), uninformative noise and large numbers of redundant features, while LLE performs slightly better at smaller effect size. Also, we applied the found methodology to large scale EHR data and found a larger separation in survival for the automatically found clusters compared to clinical classification. This is the first study to present an unsupervised homogenization pipeline designed for EHR clustering.

2.8.1 Pipeline Optimization

Normalization

EHR data are heterogeneous, containing both categorical and continuous variables at different scales. Normalization is recommended to reduce the variance among variables. Most previous studies [28, 42, 43] normalized EHR variables to a range of (0, 1), however, as shown in Table 2.2, the best normalization method is closely related to the feature reduction method. For example, for DAE and DnAE, Z-score normalization results in the best performing pipelines, while no normalization is necessary for LLE. This is reasonable since, unlike DAE and other distance-based algorithms, neighbor-based algorithms, such as LLE, eliminate the need to estimate

distance between objects.

Imputation

Given the wide array of measurements that can be obtained from patients, missing data are common, and it is impossible that every patient has every possible test and measurement. Physicians evaluate the cost-benefit of each test and may not request a particular test if the result will not be informative for the diagnosis or treatment. We evaluated a spectrum of imputation techniques that could induce different levels of artificial similarity. The simulation results favored MICE for all feature reduction methods except ISOMAP. Consistent with our studies, MICE has also shown good performance for life-history/EHR datasets in previous studies [44, 45].

The main assumptions of MICE are that other non-missing values are predictive of the missing ones (redundancy) and that the data are missing-at-random. EHR data satisfies the redundancy assumption, for example, age, sex, and height are known to be good predictors of aortic root diameter [46]. White et al. [47] note that MICE is sensitive to departures from the missing-at-random assumption. However, such assumptions can be relaxed as long as the dataset contains enough complete samples to build reliable predictive models. Theoretically, EHR data is likely to follow a missing not at random over a missing at random mechanism, as there is likely a reason for missing values (e.g. patients health, physicians recommendation, socioeconomic status). However, the true pattern of missingness is likely influenced by both MAR and MNAR. Hence, MICE can still be applied given an abundance of data.

Feature Reduction

The EHR contains many redundant pieces of information. For example, body mass index can be easily computed from height and weight. Thus, it is necessary to reduce the redundancy to extract effective (and possibly latent) features from this high dimensional dataset. Our simulation results show that among the different feature reduction methods, pipelines with DAE and LLE show the highest accuracy. Moreover, LLE outperforms DAE by 0.05-0.12 at medium effect size and high uninformative noise. This suggests that LLE might be better at detecting granular phenotypes that have more overlapped samples (1-5%, corresponding to an effect size of 4-5). Additionally, another benefit of using LLE is that no normalization to input data is needed, as discussed above.

DAE is computationally more efficient at $O(nm)$ where n is the number of samples and m is the number of features. Here, we note that LLE requires $O(m \log(k) \cdot n \log(n))$, where k denotes the number of neighbors for LLE. Once the network is trained, the weights can be applied to a new dataset with minimal computation, while LLE computes and sorts distances to all neighbors. Thus, considering the large-scale nature of the EHR data, DAE might be a better choice when used to make predictions for future patients. Recent studies deep auto-encoders have demonstrated their ability to identify meaningful representations of EHR data [42, 43]. Miotto et al. first proposed the use of deep autoencoders for EHR data and called its representation “Deep patient” [42]. They demonstrated its utility by assessing the probability of patients developing various diseases and showing improvement in classification scores for 76,214 patients and 78 different diseases. Similarly, Beaulieu-Jones et al. reported improved classification scores for amyotrophic lateral sclerosis diagnosis in clinical trials using 10,723 patients [43]. These are promising results which demonstrate the potential of the proposed pipeline with DAE to utilize EHR data to identify granular disease phenotypes, and to ultimately facilitate precise diagnoses, risk prediction

and treatment strategies. Moreover, while these previous studies have shown the promise of DAE, this is the first study to validate and design the entire pipeline for clustering. Finding novel, previously hidden features within EHR and identifying granular phenotypes from hundreds of Echocardiography measurements requires a large and comprehensive training dataset. The machine learning and clustering algorithms need to see examples of many different patients and their images in order to uncover the complex relationships that exist between their features and outcomes. The training performed on the presented dataset, with more than 400,000 studies, offers an opportunity to make precise predictions of outcomes and optimal therapies for subsequent patients. Yet, its potential is hampered by the inherent complexity and heterogeneity of EHR data.

2.8.2 EHR Application

The first complication is the missingness present in the dataset. We identify a spectrum of imputation techniques that could induce different levels of artificial similarity. Place-holders or median imputation induce the most similarity whereas predictive models, such as KNN [48] and MICE [31], induce the least similarity. The simulation results favor MICE when paired with deep-autoencoders. Moreover, MICE has also shown good performance for life-history/EHR datasets [44, 45]. Thus, we use MICE to conduct imputation.

The main assumptions of the MICE model is that other non-missing values are predictive of the missing ones (redundancy) and that the data is missing-at-random. EHR data is redundant, for example, age, gender, and height are known to be good predictors of aortic root diameter [46]. White et al. [47] note that MICE is sensitive to departures from the missing-at-random assumption. However, such assumption can be relaxed as long as the dataset contains enough complete samples to build reliable

Chapter 2. Unsupervised EHR Clustering

predictive models. Since EHR is not missing-at-random, we only keep measurements with at least 10% (40,000 studies) of valid samples and discard the rest.

The next complication in EHR data is its heterogeneity, which consists of a mixture of continuous and categorical data. We propose the use of deep-autoencoders for homogenization. Recent developments in deep auto-encoders have demonstrated their ability to identify meaningful representations of EHR data [42, 43]. Miotto et al. first proposed the use of deep autoencoders for EHR data and called its representation “Deep patient” [42]. They demonstrated its utility by assessing the probability of patients developing various diseases and showing improvement in classification scores for 76,214 patients and 78 different diseases. Similarly, Beaulieu-Jones et al. reported improved classification scores for amyotrophic lateral sclerosis diagnosis in clinical trials using 10,723 patients [43]. In contrast, we validate the practical use of the encoded representation by extracting phenotypes based on patient similarity in the compressed representation and assessing what truly matters to patients and clinicians: Are the extracted phenotypes useful for predicting outcomes such as mortality, hospitalizations, or the success of different therapies?.

The efficiency of an autoencoder can be determined by the number of encoding units necessary to reach a desired reconstruction error ($< 1\%$), see Fig. 2.5. Since the number of encoding units is typically inversely related to the reconstruction error, the optimal autoencoder is defined as the network with the least number of encoding units (minimum dimensionality) that meets the given error constraint. In that sense, autoencoders with multiple layers, deep-autoencoders, have been shown to be more efficient than shallow autoencoders [32].

2.9 Conclusion

The unsupervised deep learning analysis of EHR data from patients with HF showed superior risk stratification compared to the current paradigm of HFpEF vs HFrEF, see Fig. 2.3. The survival regression score comparison suggested a larger separation on automatically derived classes of patients. This approach may lead to more refined diagnosis and management of patients with HF.

In summary, we propose an unsupervised homogenization pipeline to fully integrate all components of EHR data for clustering patients. After MICE imputation, both LLE with raw features and DAE with z-score normalization show good clustering results. While LLE marginally outperformed DAE in several direct comparisons, the computational efficiency of DAE in evaluating new observations based on large-scale EHR data (as is desired for precision medicine approaches) provides an important advantage. Future studies are required to evaluate and compare the two pipelines in real clinical scenarios with large-scale EHR data.

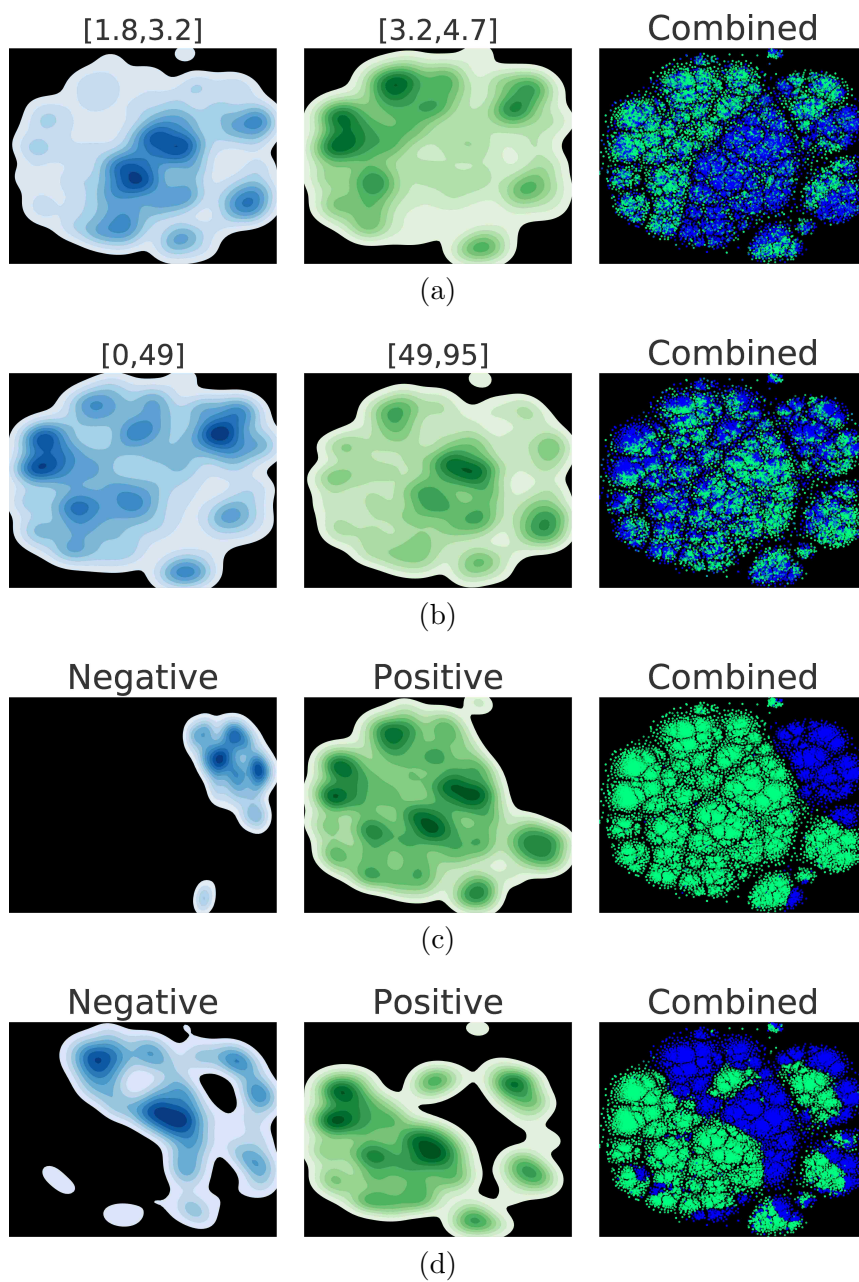


Figure 2.4: t-SNE visualization of 60 encoding units from Autoencoder. In each row, the left two plots are 2D-density plots for each category in the right plot, where each dot is a patient with Heart Failure. The same visualization is colored by (a) the Aortic Root diameter, (b) Left Ventricular Ejection Fraction, (c) Hypertension, and (d) Ischemic heart disease.

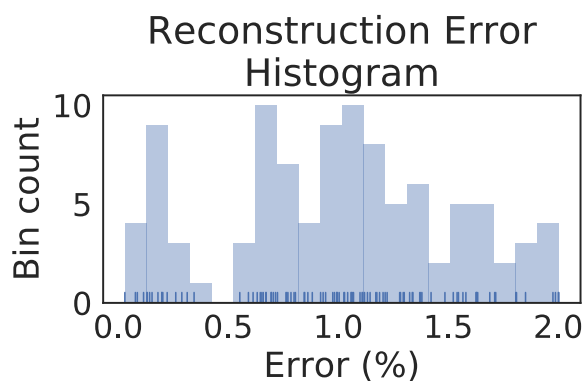


Figure 2.5: Error distribution for Autoencoder with 2 layers (500 hidden units each) and 60 encoding units

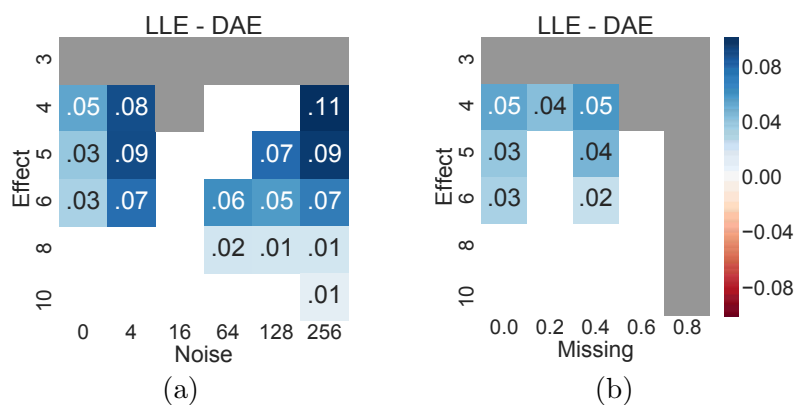


Figure 2.6: Interaction experiments for (a) Effect size vs Noise, and (b) Effect size vs Missingness. The gray areas denote where neither method scored above 0.8, the white areas denote no significant difference between score means. The colored areas denote significant differences between LLE and DAE.

Chapter 3

Echocardiography Video Processing

Imaging is critical to treatment decisions in most modern medical specialties and has also become one of the most data rich components of electronic health records (EHRs). For example, during a single routine ultrasound of the heart (an echocardiogram), approximately 10-50 videos (3,000 images) are acquired to assess heart anatomy and function. In clinical practice, a cardiologist realistically has 10-20 minutes to interpret these 3,000 images within the context of numerous other data streams such as laboratory values, vital signs, additional imaging studies (radiography, magnetic resonance imaging, nuclear imaging, computed tomography) and other diagnostics (e.g. electrocardiogram). While these numerous sources of data offer the potential for more precise and accurate clinical predictions, humans have limited capacity for data integration in decision making [49]. Hence, there is both a need and a substantial opportunity to leverage technology, such as artificial intelligence and machine learning, to manage this abundance of data and ultimately provide intelligent computer assistance to physicians [50, 51, 52, 53].

Chapter 3. Echocardiography Video Processing

Automatic video analysis has remained a challenge to date, from its transmission in clinical settings [54, 55], to its compression for a more efficient storage [56], and its use [57, 58]. An example of video analysis system for clinical use is a motion and deformation model for carotid artery plaques [59, 60], where engineered features of the plaque such as texture [61, 62, 63] were used to predict a stroke event. Another example is the detection of lesions from diabetic retinopathy [64], among others.

More recently, “deep” learning (deep neural network; DNN) technologies such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNN), Dropout Regularization, and adaptive gradient descent algorithms [12]; in conjunction with massively parallel computational hardware (graphic processing units), have enabled state-of-the-art predictive models for image, time-series, and video-based data [65, 66]. For example, DNNs have shown promise in diagnostic applications, such as diabetic retinopathy [67], skin cancer [68], pulmonary nodules [69], cerebral microhemorrhage [70, 71], and etiologies of cardiac hypertrophy [72]. Yet, the opportunities with machine learning are not limited to such diagnostic tasks [50].

Prediction of future clinical events, for example, is a natural but relatively unexplored extension of machine learning in medicine. Nearly all medical decisions rely on accurate prediction. A diagnosis is provided to patients since it helps to establish the typical future clinical course of patients with similar symptoms, and a treatment is provided as a prediction of how to positively impact that predicted future clinical course. Thus, using computer-based methods to directly predict future clinical events is an important task where computers can likely assist human interpretation due to the inherent complexity of this problem. For example, a recent article in 216,221 patients demonstrated how a Random Forest model can predict in-hospital mortality with high accuracy [18]. Deep learning models have also recently been used to predict mortality risk among hospitalized patients to assist with palliative care referrals [73]. In cardiology, variables derived from electronic health records have

Chapter 3. Echocardiography Video Processing

been used to predict two-to-five year all-cause mortality in patients undergoing coronary computed tomography [74, 75], five-year cardiovascular mortality in a general clinical population, and up to five-year all-cause mortality in patients undergoing echocardiography [24].

Notably, these initial outcome prediction studies in cardiology exclusively used human-derived, i.e. “hand-crafted” features from imaging, as opposed to automatically analyzing the raw image data. While this use of hand-crafted features is important, an approach that is unbiased by human opinions and not limited by human perception, human ability in pattern recognition, and effort may be more robust. That is, there is strong potential in an automated analysis that would leverage all available data in the images rather than a few selected clinical or clinically inspired measurements. Furthermore, the potential benefit of this approach for echocardiography may be enhanced by the added availability of rich temporal (video) data. DNNs make this unique approach possible. However, using video data also increases technical complexity and thus initial efforts to apply deep learning to echocardiography have focused on ingesting individual images rather than full videos [20].

In this chapter, we show that a DNN can predict 1-year mortality directly from echocardiographic videos with good accuracy and that this accuracy can be improved by incorporating additional clinical variables from the electronic health record. We do this through a technical advance that leverages the full echocardiographic videos to make predictions using a three-dimensional DNN. In addition to this technical advance, we demonstrate direct clinical relevance by showing that the DNN is more accurate in predicting 1-year mortality compared to two expert physician cardiologists.

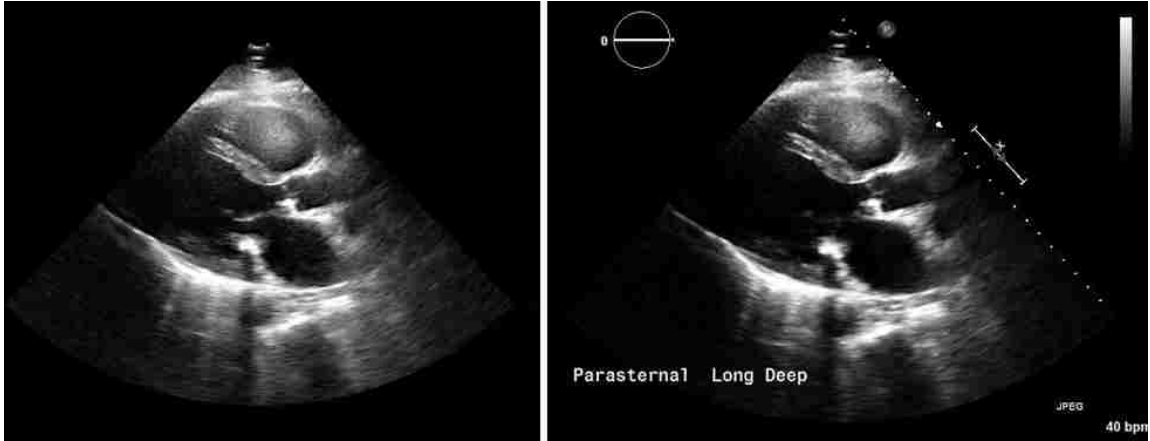


Figure 3.1: Examples of raw (left) and annotated (right) videos.

3.1 Image Collection and Preprocessing

An echocardiography study consists of several videos containing multiple views of the heart. Two clinical databases, Philips iSite and Xcelera, contained all echocardiograms collected at Geisinger. We used DCM4CHEE (version 2.0.29) and AcuoMed (version 6.0) software to retrieve a DICOM file for each echocardiography video.

The retrieved DICOM files contained an annotated video (for example, which was marked with the view name) and a raw video when the equipment was configured to store it. Without loss of generality, we used raw videos for all analyses. The raw video contained only the beam-formed ultrasound image stored in a stream of bytes format (see Figure 3.1), whereas the annotated video contained artificial annotations on top of the raw video we linearly interpolated all raw videos to 30 frames per second.

Along with the video data, the DICOM file included tags that labelled the view as to which specific image orientation was acquired. These view tags had slight variations across studies for the same type of view. For example, an apical four chamber view could be tagged as “a4”, “a4 2d”, or “ap4”. We visually inspected

samples of each unique tag and grouped them into 30 common views (see Table A.1). Since each video from a view group could potentially have different dimensions, we normalized all videos from a view to the most common row and column dimensions. We cropped/padded each frame with zeros to match the most common dimensions among the view group. We ultimately retrieved Philips-generated DICOM files with raw videos, view labels and excluded any videos that lasted less than 1 second.

3.2 Electronic health record data preprocessing

The EHR contained 594,862 echocardiogram studies from 272,280 unique patients performed over 19 years (February 1998 to September 2018). For each study, we extracted automatic and physician reported echocardiography measurements ($n = 480$) along with patient demographic ($n = 3$), vitals ($n = 5$), laboratory ($n = 2$), and billing claims data ($n = 90$; International Classification of Diseases, Tenth Revision (ICD-10), codes from patient problem lists). For measurements taken outside of the Echocardiography study, such as fasting LDL, HDL, blood pressure, heart rate, and weight and height measurements, we retrieved the closest (before or after) within a six-month window.

All continuous variables were cleaned from physiologically out of limit values, which may have been caused by input errors. In cases where no limits could be defined for a measurement, we removed extreme outliers that met two rules: 1) Value beyond the mean plus or minus three standard deviations and 2) Value below the 25th percentile minus 3 interquartile ranges or above the 75th percentile plus 3 interquartile ranges. The removed outlier values were set as missing.

We imputed the missing data from continuous variables in two steps. First, we conducted a time interpolation to fill in missing measurements using all available studies of an individual patient, i.e., missing values in between echocardiography ses-

Chapter 3. Echocardiography Video Processing

sions were linearly interpolated if complete values were found in the adjacent echocardiograms. Then, to conduct Multiple Imputation by Chained Equations (MICE) [47] and complete the entire dataset, we kept 115 of 480 echocardiography measurement variables with more than 10% non-missing measurements.

We coded the reported diastolic function in an ordinal fashion with -1 for normal, 0 for dysfunction (but no grade reported), and 1, 2 and 3 for diastolic dysfunction grades I, II, and III respectively. After imputation of the continuous measurements, we imputed the missing diastolic function assessment by training a logistic regression classifier to predict the dysfunction grade (-1, 1, 2, or 3) in a One-vs-All classifier framework using 278,160 studies where diastolic function was known.

Following imputation, we retained the physician reported left ventricular ejection fraction (LVEF) plus 57 other independent, non-redundant echocardiography measurements (i.e., excluding variables derived from other measurements; $n = 58$ echocardiography measurements in total).

We calculated the patients age and survival time from the date of the echocardiogram. The patient status (dead/alive) was based on the last known living encounter or confirmed death date, which is regularly checked against national databases in our system. We present a list and description of all 158 EHR variables used in the proposed models in the Table A.2.

3.3 Data pruning

The image collection and preprocessing resulted in 723,754 videos from 31,874 studies performed on 27,028 patients (an average of 22.7 videos per study). We linked the imaging and EHR data and discarded any imaging without EHR data. For a given survival experiment (3, 6, 9, and 12 months), we also removed studies without enough

Chapter 3. Echocardiography Video Processing

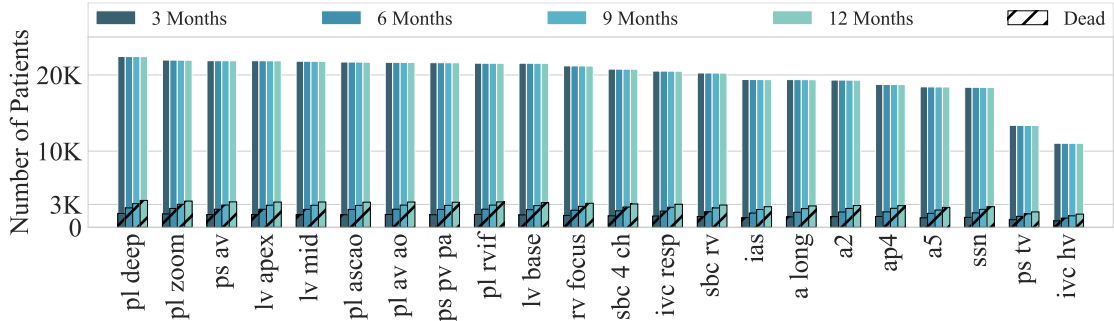


Figure 3.2: Number of patients for experiments that required 3, 6, 9, and 12 months follow-up (as indicated in the Extended Data Table 2) with the proportion of dead patients (shaded bar).

follow up. After that, we kept a single study per patient by randomly sampling one study per patient. This ensured that images from a single patient would not appear multiple times throughout training, validation, and testing groups.

We needed at least 600 patients (300 alive, 300 deceased), as indicated by a sample size calculation using the Pearson Chi-square test, to estimate and compare prognostic accuracy between the model and the two cardiologists. We assumed a 10% difference in accuracy between machine and cardiologist (80% vs 70%), 80% power, a significance level of 5%, and an approximate 40% discordancy. This was calculated using Power Analysis Software (PASS v15). Thus, we randomly sampled 300 studies of patients that survived and 300 that died within the set experiment threshold for each view, and set these aside from the valid samples to later compare the performance of the machine against two independent cardiologists. Only the parasternal long axis view (representing the best performing model and the cardiologists preference for the most comprehensive single view) was ultimately used for the cardiologist comparison. The total number of valid samples for each experiment and view is shown in Table A.3, and Figure 3.2.

We excluded parasternal long mitral valve, parasternal long pulmonic valve, short axis apex zoom, short axis mid papillary zoom, parasternal long lax, apical 3 zoom, and apical 2 zoom views, as they did not have enough available samples to run the experiments.

3.4 Model selection

For Echocardiography video classification, we explored four different architectures: 1) A time-distributed two-dimensional Convolutional Neural Network (2D CNN) with Long Short-Term Memory (LSTM), 2) a time-distributed 2D CNN with Global Average Pooling (GAP), 3) a 3D CNN and 4) a 3D CNN with GAP. For simplicity, we abbreviate the four candidate architectures: 2D CNN + LSTM, 2D CNN + GAP, 3D CNN, and 3D CNN + GAP.

The 2D CNN + LSTM consisted of a 2D CNN branch distributed to all frames of the video. This architecture was used for a video description problem [76], where all frames from a video belonged to the same scene or action. Since all frames of the echocardiography video belong to the same scene or view, it is correct to assume that the static features would be commonly found by the same 2D kernels across the video. This assumption was put in practice for echocardiography view classification [77]. The LSTM layer aggregates the CNN features over time to output a vector that represents the entire sequence.

The 2D CNN + GAP approach exchanged the LSTM layers for the average CNN features as a time aggregation of frames. The GAP layer provides two advantages. It requires no trainable parameters, saving 1008 parameters from the LSTM layers, and enables feature interpretation. The final fully connected layer after the GAP would provide a weighted average of the CNN features, which could indicate what sections of the video weighted more in the final decision. The 3D CNN approach

Chapter 3. Echocardiography Video Processing

aggregates time and space features as the input data flows through the network.

3D CNNs have also shown successful applications for video classification. As opposed to the 2D CNN approach, 3D CNN incorporates information from adjacent frames at every layer, extracting time-space dependent features.

The 3D CNN approach would replace the Flatten operation for a GAP layer. In a similar fashion to the 2D CNN + GAP approach, the GAP layer would reduce the number of input features to the final Dense layer, thus the reduction of the number of parameters from 641 to 17; while enabling the traceback of the contributions of video features.

We defined the convolutional units of the 2D and 3D CNNs as a sequence of 7 layers in the following composition: CNN layer, Batch Normalization, ReLU, CNN layer, Batch Normalization, ReLU, and Max Pooling (see Figure 3.3). All kernel dimensions were set to 3 and Max Pooling was applied in a 3 x 3 window for 2D kernels and 3 x 3 x 3 for 3D kernels.

A detailed description of the number of parameters for the 2D CNN + LSTM architecture is shown in Table 3.2, 2D CNN + GAP is shown in Table 3.3, 3D CNN is shown in Table 3.4, and 3D CNN + GAP is shown in Table 3.5. We applied all four candidate architectures to all the identified echocardiography views with a 1-year mortality label, and the 3D CNN showed consistently the best performance, see Figure 3.4.

Similarly, we assessed the performance gain at different image resolutions. We reduced the video resolution by factors of 2, 3, and 4. No consistent significant loss in performance was observed across all views, see Figure 3.5. Thus, we decided to conduct all experiments with a resolution reduction by a factor of 4 to reduce computational cost. To incorporate EHR data into the prediction, we trained a three-layer multi-layer perceptron (MLP) with 10 hidden units at each layer. Then,

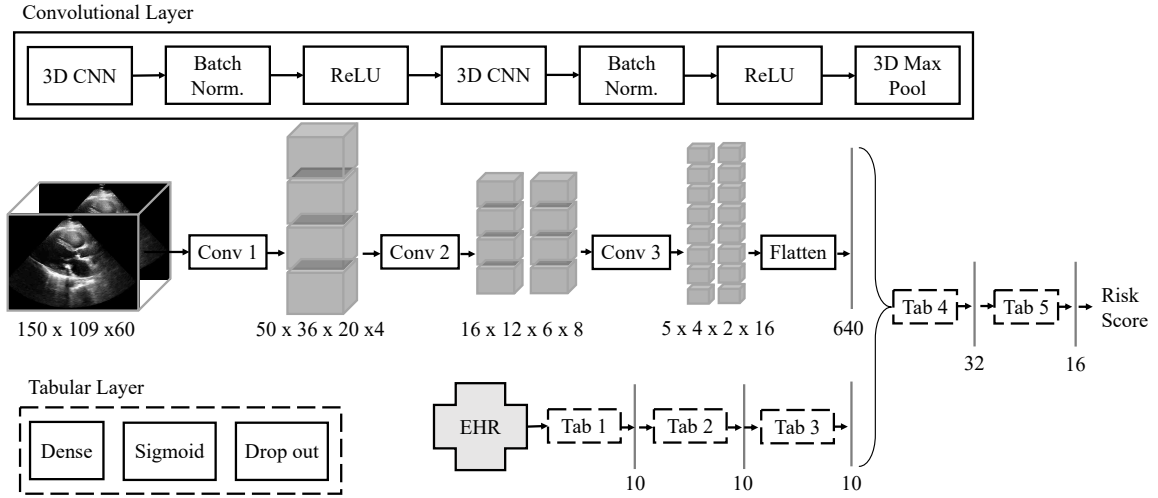


Figure 3.3: Neural network architecture for mortality prediction from echocardiography videos and electronic health record (EHR) data. The convolutional layer (Conv) is shown on the top box with a solid outline and the tabular layer (Tab) is shown in the bottom box with a dashed outline. The convolutional layer consists of Convolutional Neural Networks (CNN), Batch Normalizations (Batch Norm.), rectified linear units (ReLU), and a three-dimensional Maximum Pooling layer (3D Max Pool). The tabular layer consists of a fully connected layer (Dense) with sigmoid activations and a Drop Out layer. The input video dimensions were $150 \times 109 \times 60$ pixels, and the output dimension of every layer are shown.

we concatenated the last 10 hidden units with the CNN branch, see Figure 3.3.

3.4.1 Effect of adding optical flow inputs

Optical flow velocity maps have been shown to be informative along with the original videos for classification tasks [78]. Thus, we computed the dense optical flow vectors of the echocardiography raw videos using the Gunnar Farnebacks algorithm as implemented in the OpenCV (version 2.4.13.7) software library. We set the pyramid scale to 0.5, the number of levels to 3, and the window size to 5 pixels. The vectors were then converted to color videos where the color indicated direction (as in the HSV color space) and the brightness denoted amplitude. This resulted in an

Chapter 3. Echocardiography Video Processing

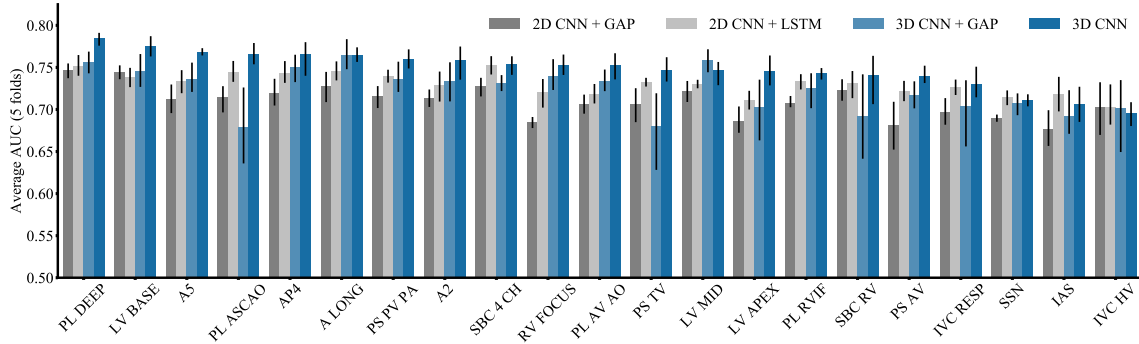


Figure 3.4: AUCs of one-year mortality predictions across all views with four different neural network architectures: 2D CNN + Global Average Pooling (GAP; dark gray), 2D CNN + Long Short-Term Memory (LSTM; light gray), a 3D CNN + GAP (light blue), and 3D CNN (dark blue).

image video that was fed to the neural network model through an independent 3D CNN branch along with the raw video. As seen in Figure 3.6, this combination of the optical flow video to the raw video did not yield consistently improved model performance compared with models using the raw video alone. Therefore, we did not use optical flow for the final study analyses.

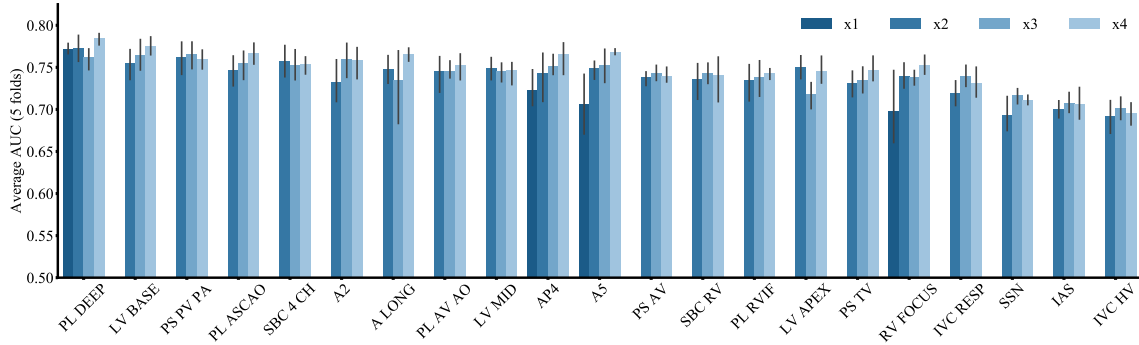


Figure 3.5: AUCs of one-year mortality predictions across all views with different levels of reduced resolution ranging from native (x1) to 4-fold (x4). Note that full native resolution training was only done for select views due to the computational time required to complete the experiment at this resolution.

Chapter 3. Echocardiography Video Processing

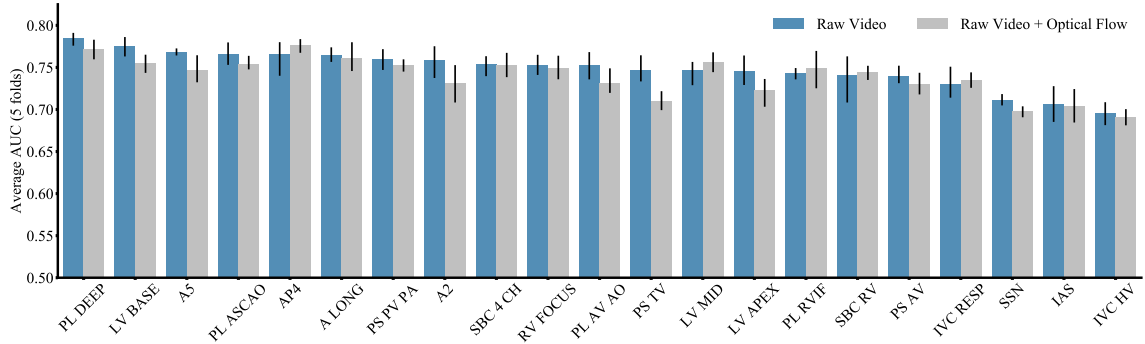


Figure 3.6: One-year mortality prediction performance ranking for all echocardiography views using only the raw video (blue) versus the raw video with optical flow features (gray).

3.5 Training Procedure

We used the RMSProp [79] algorithm to train the networks with LSTM coupling, and AdaGrad [14] for the 3D CNN architectures. Each iteration of the 5-fold cross validation contained a training, validation, and test set. The training and test sets were sampled such that they had the same prevalence of alive patients, but the validation set was sampled with a balanced proportion. The validation set comprised 10% of the training set.

As we trained the DNN, we evaluated the loss (binary cross-entropy) on the validation set at each epoch. If the validation loss did not decrease for more than 10 epochs we stopped the training and reported the performance, in AUC, of the test set. We set the maximum number of epochs to 1000 and kept the default training parameters as defined by the software Keras (version 2.2). Training always ended before the maximum number of epochs was reached.

Since the prevalence of each patient class is imbalanced (16% deceased patients),

we set the weights for each class as follows:

$$w_i = \frac{\text{Total Number of Samples}}{2(\text{Number of Samples in class } i)}$$

All training was performed in an NVIDIA DGX1 platform. We independently fit each fold on each of the 8 available GPUs. The main experiment, shown in Figure 3.4, took a total of six days to complete.

3.6 Cardiologist Survey Dataset

We set a 600-patient survey used to compare the accuracies of the cardiologists and the model, as described in the data pruning section, was intentionally balanced with respect to mortality outcomes (300 dead and 300 alive at one year) in order to ensure adequate power to detect differences in performance. The cardiologists were blinded to this distribution at the time of the review. We note that this balance is not reflective of typical clinical outcomes, particularly in a primary or secondary care setting, in which the base rate for 1-year survival is much higher. Hence, we cannot claim that this survey comparison between cardiologists and the model, as implemented, represents prediction in a realistic clinical setting. We do note, however, that the realistic clinical survival base rate was represented in the model training/testing sets, just as in the conditioning experiences of the cardiologists (consistent with their preference-high specificity for death in over-estimating 1-year survival). Thus, the model was not advantaged in this regard by learning to expect this different outcome. Instead, rather than prediction informed by clinical base rates, our comparison sought to evaluate the true discriminative abilities and accuracies of the cardiologists compared to the machine.

Chapter 3. Echocardiography Video Processing



Figure 3.7: Interface of the web application developed for cardiologists to predict survival one year after echocardiography

3.6.1 Software for cardiologist survey

We deployed a web application with the interface shown in Figure 3.7. The application required the cardiologist to input their institutional credentials for access. We showed the 10 EHR variables and the two versions of the video, raw and annotated. The application then recorded the cardiologist prediction as they clicked on either the “Alive” or “Dead” buttons.

3.6.2 Statistical analysis: Machine vs Cardiologists

The cardiologists responses were binary, and the Machines response was continuous. We set 0.5 as the threshold for the Machines response prior to performing the final comparison experiment. Since all responses were recorded for the same samples, we conducted a Cochran's Q test to assess whether the three responses were significantly different in the proportion of correctly classified samples.

3.7 Results and Discussion

Ultimately, we utilize a fully 3D Convolutional Neural Network (CNN) design in this study. CNNs are neural networks that exploit spatial coherence in an image to significantly reduce the number of parameters that a fully connected network would need to learn. CNNs have shown promise in image classification tasks [12], even surpassing human abilities [80]. Details of additional model architectures attempted (including a time-distributed 2D CNN + long short term memory network [LSTM] [81, 82, 83, 84]) are described in the methods.

We first collected 723,754 clinically acquired echocardiographic videos (approximately 45 million images) from 27,028 patients that were linked to at least 1 year of longitudinal follow-up data to know whether the patient was alive or dead within that time frame. Overall, 16% of patients in this cohort were deceased within a year after the echocardiogram was acquired. Based on a power calculation detailed in the methods, we separated data from 600 patients for validation and comparison against two independent cardiologists and used the remaining data for 5-fold cross-validation schemes.

During the acquisition of an echocardiogram, images of the heart and large blood vessels are acquired in different two-dimensional planes, or “views”, that are stan-

Chapter 3. Echocardiography Video Processing

standardized according to clinical guidelines [85]. We generated separate models for each of the 21 standard echocardiographic views and showed that the proposed models were able to accurately predict 1-year survival using only the raw video data as inputs (Figure 3.4). The chosen 3D CNN architecture (AUC range: 0.695-0.784) outperformed the 2D CNN + LSTM architecture (AUC range: 0.703-0.752) for most views. In both cases, the parasternal long-axis (“PL DEEP”) view had the best performance. This result was in line with clinical intuition, since the PL DEEP view is typically reported by cardiologists as the most informative “summary” view of overall cardiac health. This is because the PL DEEP view contains elements of the left ventricle, left atrium, right ventricle, aortic and mitral valves, and whether or not there is a pericardial or left pleural effusion all within a single view.

These results were relatively insensitive to image resolution (no significant difference was observed between models using full native resolution images (400 x 600 pixels) and reduced resolution images (100 x 150 pixels); see Figure 3.5). Similarly, adding derived optical flow velocity maps [78] to the models along with the pixel level data did not improve prediction accuracy, see Figure 3.6.

Next, we investigated the predictive accuracy of the models at additional survival intervals, including 3, 6, 9, and 12-month intervals after echocardiography. The models generally performed better at longer intervals, but AUCs for all cases were greater than 0.64 (Figure 3.8).

We then added select clinical (“EHR”) variables from each patient including age, tricuspid regurgitation maximum velocity, heart rate, low density lipoprotein [LDL], left ventricular ejection fraction, diastolic pressure, pulmonary artery acceleration time, systolic pressure, pulmonary artery acceleration slope, and diastolic function. These 10 variables have previously been shown to contain >95% of the power for predicting 1-year survival in 171,510 patients [24] and their addition improved accuracy to predict 1-year survival for all echocardiographic views, with AUCs ranging

Chapter 3. Echocardiography Video Processing

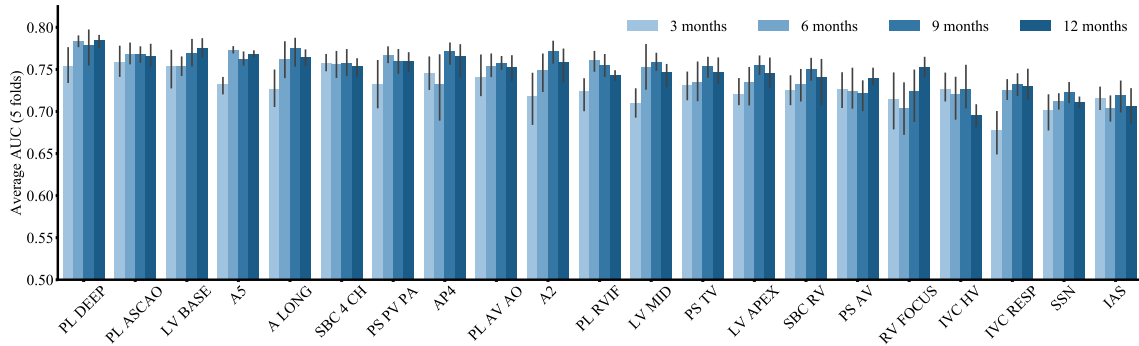


Figure 3.8: Mortality prediction performance for echocardiographic videos alone at 3, 6, 9 and 12 months for all views. The error bars denote one standard deviation above and below the average across 5 folds.

from 0.79-0.82 (compared to 0.70-0.78 without these 10 EHR variables).

Next, we developed a software platform (see section 3.6.1) that we used to display an echocardiographic video of interest along with the 10 select EHR variables to two independent cardiologist echocardiographers who were blinded to the clinical outcomes. The cardiologists assessed whether each of 600 patients (independent test set extracted randomly from the original dataset of parasternal long axis views and not used for training of the machine) would be alive at one year based on the data presented. The final trained model (trained in all but these 600) was also applied to the same independent test set.

The overall accuracy of the model (75%) was significantly higher than that of the cardiologists (56% and 61%, $p = 4.2 \times 10^{-11}$ and 6.9×10^{-7} by Bonferroni-adjusted post-hoc analysis, Figure 3.9a). We found that the cardiologists tended to overestimate survival likelihood, yielding high specificities (97% and 91%, respectively) but poor sensitivities (16% and 31%, respectively) while the model, by design, balanced sensitivity and specificity (both 75%). Moreover, as demonstrated in Figure 3.9b, the operating points for the individual cardiologists fell within the envelope of the model’s receiver operating characteristic curve (as opposed to falling at a different

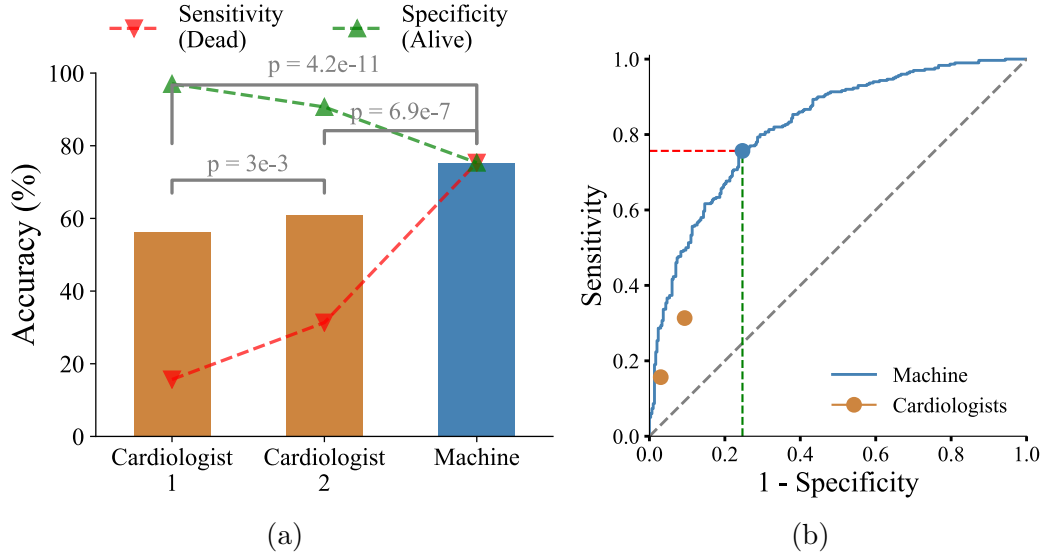


Figure 3.9: Cardiologists vs Machine performance for 1-year mortality prediction from the survey dataset of 600 samples with balanced prevalence. The left plot (a) shows the accuracy in bars and sensitivity (red) and specificity (green) as triangles. The right plot (b) shows the operating points of the cardiologists as orange dots, the Receiver Operating Characteristic curve for the machine performance in blue, and the machine operating point as a blue dot.

point on the same curve), suggesting inferior predictive performance in this task.

Beyond the limited inputs selected for the clinical expert comparison, we sought to further characterize the model performance unconstrained by data input limitations. That is, we completed additional experiments permuting the input combinations of structured data (none, limited set [top 10 EHR variables], full set [158 EHR variables, as described in methods]) and echocardiography videos (none, single view, all 21 views). Models without videos were trained using all available data in our structured echocardiography measurement database (501,449 valid studies), while the models with videos were trained with all videos available for each view, ranging from 11,020 to 22,407 for single videos and 26,428 combined. In all cases, the test set was the 600 patients held out for the clinical expert comparison.

Table 3.1: AUC scores for each data modality combination of EHR and Echo video data on the 600 left out studies used to compare to the cardiologists. No video models were trained on all available studies, whereas Single Video and All Videos were trained on a subset where video data were available. The No EHR variables and No Video cell denotes a random guess.

	No Video (~500K samples)	Single Video (~22K samples)	All Videos (~27K samples)
No EHR	0.532	0.801	0.839
Limited EHR	0.786	0.824	0.843
Full EHR	0.851	0.825	0.858

Table 3.1 shows that all videos combined with the full EHR variable set had the highest AUC in the held out test set of 600 studies, demonstrating the potential to further enhance the performance of the already clinically superior model. Several general trends were also noted. First, a single video view out-performed a model that included 10 EHR variables as input. Second, multiple videos had higher performance than single videos. Third, the learning curves (Figure 3.10) for multi-video predictions demonstrated that, despite having access to a massive dataset (26,428 echocardiographic videos), more samples would likely result in even higher performance for multi-video predictions. In contrast, the performance of the full EHR data-only model, which was consistently less than the full EHR plus videos model, was beginning to plateau. Hence, our novel multi-modal DNN approach, inclusive of echocardiography videos, provides enhanced performance for this clinical prediction task compared to what can be achieved using EHR data alone (inclusive of hand-crafted features derived by humans from the videos).

3.8 Conclusion

Here we demonstrated the potential for DNNs to help cardiologists predict a clinically relevant endpoint, mortality after echocardiography, using both raw video data and

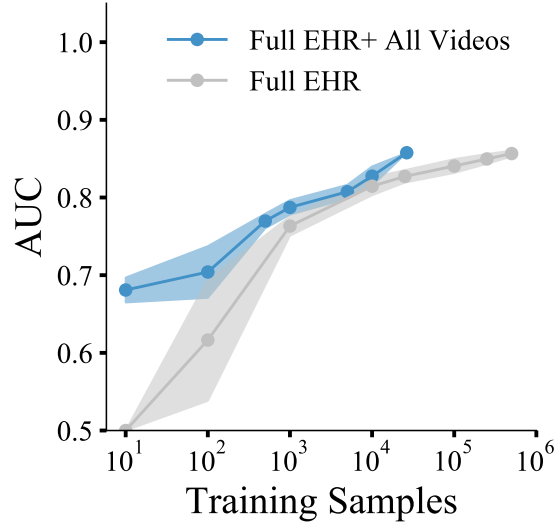


Figure 3.10: Learning curves for the full (158) EHR variables model compared to the full EHR variables plus videos. The AUC is reported on the 600 patient set as a function of training set size, ranging from 10 to the maximum number of datasets available for the given data inputs, which was 501,449 for the EHR variables and 26,428 for the Full EHR+videos.

relevant clinical data extracted from the electronic health record. For training the DNN, we leveraged a massive dataset of 723,754 clinically-acquired videos of the heart consisting of 45 million images. We showed that the ability of our DNN to discriminate 1-year survival even with limited model inputs surpassed that of trained cardiologists, suggesting that these models can add value beyond a standard clinical interpretation. To our knowledge, no prior study has demonstrated the ability to train a deep neural network to predict a future clinically-relevant event directly from image pixel-level data. Additional experiments demonstrated opportunities to achieve further significant performance gains by incorporating more EHR variables, simultaneously using all echocardiography views, and leveraging more data for model training.

We chose 1-year all-cause mortality as a highly important, easily measured clin-

ical outcome to demonstrate feasibility for this initial work. Importantly, all-cause mortality is a well-defined endpoint without the bias that can be introduced into endpoints such as cardiovascular-specific mortality, and it can easily be extracted from an EHR that is validated against national death index databases. Moreover, mortality prediction is highly relevant for numerous applications in cardiology, as evidenced by the multitude of clinical risk scores that are currently used clinically (Framingham [7], TIMI [86], and GRACE [87] scores, etc).

3.9 Future work

Future research will be needed to evaluate the performance of these models to predict additional clinically relevant outcomes in cardiology, such as hospitalizations or the need for major procedures such as a valve replacement.

Though these data had inherent heterogeneity since they were derived from a large regional healthcare system with over 10 hospitals and hundreds of clinics, additional data from other independent healthcare systems will be required to assess generalizability. Future work should be able to further improve accuracy by combining multiple videos into a single model, including Doppler based videos. Thus, methodology and architecture have been developed while feasibility and significant potential have been demonstrated for extracting predictive information from medical videos. With the ongoing rate of technological advancement and the rapid growth in electronic clinical datasets available for training, neural networks will augment future medical image interpretations with accurate predictions of clinical outcomes.

Table 3.2: Low-parameter 2D CNN + LSTM with 4,237 trainable parameters

Layer	# Parameters	Description
Input	-	109x150x60 video
L1: Conv1+ReLU	40	4 2D feature maps
L1: Conv2+ReLU	148	4 feature map groups
L1: Batch norm.	8	Normalize feature maps
L1: Max Pool	-	3x3 max-pooling
L2: Conv3+ReLU	296	8 feature map groups
L2: Conv4+ReLU	584	8 feature map groups
L2: Batch norm.	16	
L2: Max Pool	-	3x3 max-pooling
L3: Conv5+ReLU	584	8 feature map groups
L3: Conv6+ReLU	584	8 feature map groups
L3: Batch norm.	16	
L3: Max Pool	-	3x3 max-pooling
L4: Conv7+ReLU	584	8 feature map groups
L4: Conv8+ReLU	584	8 feature map groups
L4: Batch norm.	16	
L4: Max Pool	-	3x3 max-pooling
L5: LSTM	544	8 hidden units
L6: LSTM	208	4 hidden units
Dense+ReLU	20	4 hidden units
Output+Sigmoid	5	1 output unit

Table 3.3: Low-parameter 2D CNN + GAP with 3,477 trainable parameters

Layer	# Parameters	Description
Input	-	109x150x60 video
L1: Conv1+ReLU	40	4 2D feature maps
L1: Conv2+ReLU	148	4 feature map groups
L1: Batch norm.	8	Normalize feature maps
L1: Max Pool	-	3x3 max-pooling
L2: Conv3+ReLU	296	8 feature map groups
L2: Conv4+ReLU	584	8 feature map groups
L2: Batch norm.	16	
L2: Max Pool	-	3x3 max-pooling
L3: Conv5+ReLU	584	8 feature map groups
L3: Conv6+ReLU	584	8 feature map groups
L3: Batch norm.	16	
L3: Max Pool	-	3x3 max-pooling
L4: Conv7+ReLU	584	8 feature map groups
L4: Conv8+ReLU	584	8 feature map groups
L4: Batch norm.	16	
L4: Max Pool	-	3x3 max-pooling
L5: GAP	0	Global average Pooling
Output+Sigmoid	17	1 output unit

Table 3.4: Low-parameter Dyadic 3D CNN with 14,309 trainable parameters. For the dimension of the 3D CNN architecture, refer to Fig. 4.2.

Layer	# Parameters	Description
Input	-	60x109x150 video
L1: 3D CNN 1	112	4 3D feature maps
L1: Batch norm.	8	Normalize feature maps
L1: 3D CNN 2	436	4 3D feature maps
L1: Batch norm.	8	Normalize feature maps
L1: Max Pool	-	3x3x3 max-pooling
L2: 3D CNN 3	872	8 3D feature maps
L2: Batch norm.	16	Normalize feature maps
L2: 3D CNN 4	1,736	8 3D feature maps
L2: Batch norm.	16	Normalize feature maps
L2: Max Pool	-	3x3x3 max-pooling
L3: 3D CNN 5	3,472	16 3D feature maps
L3: Batch norm.	32	Normalize feature maps
L3: 3D CNN 6	6,928	16 3D feature maps
L3: Batch norm.	32	Normalize feature maps
L3: Max Pool	-	3x3x3 max-pooling
Flatten	-	Vectorization
Output+Sigmoid	641	1 output unit

Table 3.5: Low-parameter Dyadic 3D CNN + GAP with 13,685 trainable parameters. For the dimension of the 3D CNN architecture, refer to Fig. 4.2.

Layer	# Parameters	Description
Input	-	60x109x150 video
L1: 3D CNN 1	112	4 3D feature maps
L1: Batch norm.	8	Normalize feature maps
L1: 3D CNN 2	436	4 3D feature maps
L1: Batch norm.	8	Normalize feature maps
L1: Max Pool	-	3x3x3 max-pooling
L2: 3D CNN 3	872	8 3D feature maps
L2: Batch norm.	16	Normalize feature maps
L2: 3D CNN 4	1,736	8 3D feature maps
L2: Batch norm.	16	Normalize feature maps
L2: Max Pool	-	3x3x3 max-pooling
L3: 3D CNN 5	3,472	16 3D feature maps
L3: Batch norm.	32	Normalize feature maps
L3: 3D CNN 6	6,928	16 3D feature maps
L3: Batch norm.	32	Normalize feature maps
L3: Max Pool	-	3x3x3 max-pooling
L4: GAP	-	Global Average Pooling
Output+Sigmoid	17	1 output unit

Chapter 4

Multimodal Interpretable Risk Prediction

The adoption of Electronic Health Records (EHR) in medicine has facilitated the collection of massive amounts of clinical data which can be used to develop highly accurate risk models that physicians can use to guide medical decision making. To take full advantage of the available EHR data, these models, similar to a physician, need to be able to handle multiple modalities as inputs and explain its decisions. For example, both tabular data such as laboratory measurements and pixel data from clinical images should be readily incorporated. This basic framework is shown in Fig. 4.1.

As documented in [24, 88, 89, 90], precision medicine can benefit greatly from development of these risk models. The proliferation of these models has prompted scrutiny from the medical community, which demands clinical validity and interpretability to improve usefulness [91] and inclusion of all relevant predictors (or, conversely, explanation when a relevant data input is excluded) [92]. Moreover, the recent European General Data Protection Regulation (<https://eugdpr.org/>) states

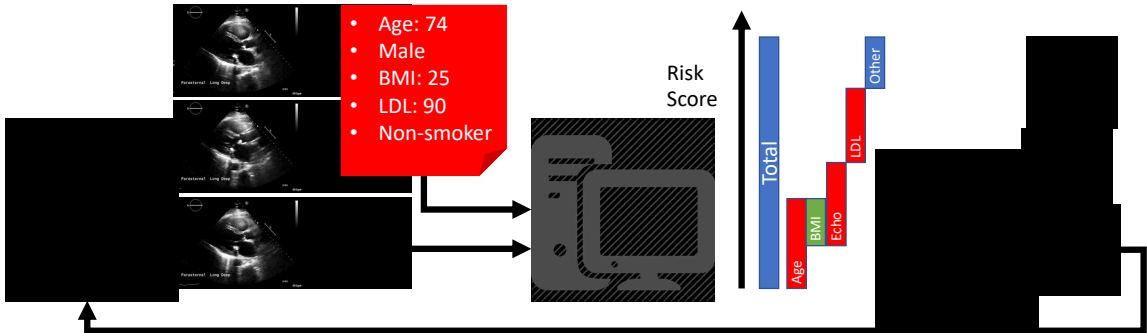


Figure 4.1: General framework for multi-modal risk assessment. EHR data and cardiac ultrasound videos are input to the risk assessment system. We emphasize the use of separable non-linear models where we look at contributions from each modality and each feature separately and also within the joint multi-modal framework. The mortality risk assessment is used to inform treatment.

that individuals who have decisions made about them by algorithms have a right to know the basis of the decision and the factors that influenced this decision. Thus, any medical risk model should be interpretable and facilitate understanding of the various contributions of different inputs towards the overall risk assessment.

When only using tabular EHR data, clinical interpretability is well supported by linear models. Some examples of clinical adoption of linear models are the Framingham risk score [23], which yields a score for the risk of developing a cardiovascular disease within ten years, and the Seattle Heart Failure score [8], which predicts 1-, 2-, and 3-year mortality in patients with Heart Failure.

The coefficients of a multivariate linear predictor can be used to assess feature importance based on its magnitude, and the effect directionality based on its sign. Unfortunately, the performance of linear models can be limited. As described in [24], non-linear models such as random forests outperform linear models for predicting mortality risk using EHR data. While not as direct as linear models, there are also approaches to support explanations for non-linear methods. As an example, for ensemble methods based on decision trees (e.g., Random Forests), we can rank the

input features based on the proportion of samples that appear at each decision node where each feature is used. Currently, methods to support clinical interpretability include building a single tree with multivariate decision nodes [93], extracting an optimal tree with a minimum performance cost [94], and an indirect method that offers recommendations for transforming true negative instances into positively predicted ones [95]. However, the ability to expand such interpretable models to more robust, multi-modal frameworks—capable of ingesting all the diverse and heterogeneous elements of EHR data, such as digital images and videos—has been challenging. To date, no such model has been developed.

A major challenge in developing interpretable multi-modal models for clinical use is that non-interpretable deep learning methods dominate research on data such as medical images [16, 67, 96] and tabular EHR data [17]. To particularly highlight the capabilities of deep learning in medical imaging analysis, such methods have been used in ultrasound video analysis for frame labeling tasks such as segmentation of certain chambers of the heart (the left ventricle) [19, 20], fetal standard image plane and orientation detection [21, 22], and echocardiography video classification tasks, as in chapter 3. Given this success, there is clear need to explore/develop interpretable frameworks that are compatible with deep learning models.

4.1 Low Number of Parameters Networks

The rapid rise of deep learning methods has also been associated with the development of lower-parameter neural network systems that can also deliver better performance than previously considered methods. To demonstrate the trend, we consider some of the most popular successful classifiers. In 2012, AlexNet used 60M parameters to achieve a top-5 test error rate of 15.3% for the ILSVRC-2012 competition [2]. In 2016, the updated version of the Inception architecture used about 25M param-

Table 4.1: Comparison of Parameters/Cases ratio of different ImageNet models: (i) AlexNet [2], (ii) Inception V3 [3], (iii) DenseNet [4].

Deep Learning Network	Pars/Cases	Percentage Ratio
AlexNet	60M/1.2M	5,000%
Inception V3	25M/1.3M	1,923%
DenseNet-201	20M/1.3M	1,538%
Low-Comp. 2D CNN	4,237/31,278	14%
Low-Comp. 3D CNN	14,309/31,278	46%

ters to achieve a top-5 test error rate of 5.6% for the same competition [3]. In 2017, DenseNet-201 used 20M parameters model to achieve 6.34% accuracy on the same dataset and thus match the performance of a 101-layer ResNet with more than 40M parameters [4].

We introduce low-parameter convolutional neural network architectures with a small number of layers to process cardiac ultrasound videos. To keep the number of parameters low, for the 2D CNN, we stacked two LSTM output layers, as suggested in [97], for capturing the temporal dependencies in the data. Then, for the 3D CNN, we introduce a dyadic approach where the number of 3D feature maps doubles for each ConvNet (4, 8, 16). As a result, to recognize the low number of parameters that we are considering here, we note that the number of parameters divided by the number of cases is just 14% for our 2D CNN and 44% for our 3D CNN. In comparison, for the same ratio, AlexNet is at 5,000%, the latest inception model is at 1,900% that only drops to 1,538% for DenseNet-201 (see Table 4.1). Furthermore, the proposed CNN architectures use a small number of layers to support better interpretability, since the complexity of interpreting the feature maps increases with the number of filters and layers.

The proposed CNN architectures represent optimal representations that were obtained through extensive experimentation. We refer to chapter 3 for details of our approach. More specifically, we investigated the use of different image resolutions,

the addition of optical flow feature maps, and alternative echocardiography video views. We have found the parasternal long axis view to be optimal.

4.2 Interpretability and Explainability

Interpretation of deep learning models remains a challenge. Some efforts to interpret deep learning models, documented in section VI of [17], are maximum activation, imposing constraints, qualitative clustering, and a mimic learning method that approaches deep learning performance using a gradient boosting tree. Unfortunately, maximum activation is impractical for global interpretations since there is a very large number of internal neurons that can be maximized. Imposing constraints can limit the search space and help interpretability. As we shall see, we will also impose non-negativity constraints to resolve ambiguities.

Early efforts to provide feature importance in neural network models have been reported by Gevrey et. al in [98]. In classical stepwise selection, feature importance is assessed based on performance changes. More directly, we can use the partial derivative of the output with respect to a specific input feature to assess their linear dependency. In this case, a positive partial derivative implies that an increase in the input feature value will also result in an increase in the output. On the other hand, a negative partial derivative indicates that an increase in the feature will result in a reduction of the output.

For image analysis applications, two additional approaches have been introduced. First, for convolutional neural network (CNN) based methods, we can look at the output images from each layer to understand how the CNN performs feature extraction at different levels. Second, more recently, there is an effort to explain CNNs that perform semantic segmentation. The basic approach described in Fully Convolutional Networks (FCN) [99], SegNet [100], and U-net [101], is to use an auto-encoding

Chapter 4. Multimodal Interpretable Risk Prediction

structure that predicts class labels of each pixel by using a transposed version of a traditional CNN architecture. Yet, large scale semantic labeling of big datasets, such as those available in the EHR, is intractable.

Other approaches focus on building around a non-interpretable model to provide explanations. Local Interpretable Model Explanations (LIME) [102], builds local interpretable models that capture the behavior of the network for small variations of a given input and provides feature ranking by presenting the coefficients of an approximated linear model. The Anchor framework [103, 104] improves the precision of LIME with if-then rules that represent local sufficient conditions for the network to make the prediction. Model Agnostic Globally Interpretable Explanations (MAGIX) [105] LIME for global explanations also in the form of if-then rules.

Both the partial derivative and the LIME approaches cannot provide global descriptions of the input effects. To understand this problem, we note that local linear models can vary significantly from sample to sample. Hence, when using LIME or partial derivatives, there is a need to specify all of the inputs and then fit the local linear model to the specific patient.

MAGIX addresses the issue of global explainability, but it requires categorization of all input variables into bins with pre-defined cut-off intervals. Unfortunately, by binning the input variables, MAGIX may cause a classifier to lose granularity and accuracy potential. Furthermore, it is clear that patient response to disease progression is best modeled using continuously varying input parameters. In our proposed approach, we model both continuous and large-scale effects.

While explanations are a step forward to interpretability, inherently interpretable models are still desired in high stakes decisions such as the prediction of clinical outcomes [25]. Interpretable models can be based on applying or extending classical models, such as Generalized Additive Models [106]. The basic idea is to design in-

Chapter 4. Multimodal Interpretable Risk Prediction

herently interpretable models which once trained, yield inter-modality feature importance, feature response functions, and intuitive interpretations. Logistic regression is the standard example in this category. We extend this classical logistic regression approach to a multi-modal framework with polynomial transformations on continuous input variables and multi-modal training. The coefficients of the logistic regression, the polynomial coefficients, and parameters from other modalities are all trained concurrently. Here, we acknowledge that the use of sigmoid activation functions in deep learning systems also represents an extension of logistic regression. However, the use of a large number of layers in deep learning systems makes interpretability impossible. In contrast, our proposed approach remains fully interpretable and leads to dramatically improved performance rates with a small number of parameters. In fact, as we shall show in our results, our proposed approach out-performs logistic regression and approximates the performance of advanced non-linear models (Random Forests and XGBoost) by achieving better performance at small number of input features. Moreover, our approach enables us to rank features across different modalities (e.g., time-series, image and video, and tabular data).

Our basic approach is to consider polynomial transformations of each scalar input factor separately and then use a simple weighted sum to combine their contributions (along with other inputs, including video, binary, or continuous variables) for predicting mortality risk. This approach has several advantages. First, we can use the weights to assess the importance of each scalar factor. Second, we can provide an independent global assessment of the contribution of each scalar factor. Third and most importantly, by building models based on the ranked factors, the proposed approach can achieve excellent classification performance with just a small number of factors. Furthermore, in terms of classification performance, the proposed approach out-performs linear regression while approximating the performance of non-linear and non-interpretable approaches.

4.3 Experimental Setup

Here, we demonstrate the value of this approach by applying it to several different types of multi-modal input datasets with the goal of predicting the risk of 1-year mortality after echocardiography. We utilize three different sets of input variables: (i) clinical data (CD) only (e.g. age, sex, diagnoses and laboratory values) (ii) numeric variables derived from echocardiography videos, which we call echocardiography video measurements (EVM), and (iii) Echocardiography Video (EV), that is pixel data from the parasternal long axis view. By considering models that utilize different variable inputs, we can investigate the contributions of each modality separately. For example, by comparing predictions derived using EVM only against the results from video analysis, we establish that EV is more effective than EVM for risk assessment, even with a single video out of more than 20 that are typically acquired during a session and used to derive the “EVM” inputs). On the other hand, we also establish that risk assessment based on multiple modalities significantly outperforms predictions based on any single modality. Furthermore, multi-modality feature ranking provides a quantitative assessment of how features from each modality contribute to optimal risk assessment.

4.4 Interpretable Neural Network

We present the overall architecture of the proposed model in Fig. 4.2, with 100 scalar variables from clinical data (CD), 58 Echocardiography Video Measurements (EVM), measured from the video data by clinicians or technologists, and a Echocardiography Video (EV) from the parasternal long axis view. The model infers from the input data to produce a risk score that represents the likelihood of mortality within a year of the echocardiography study.

Chapter 4. Multimodal Interpretable Risk Prediction

In order to integrate clinical features from multiple modalities, we differentiate between categorical factors (e.g., sex), continuous clinical factors (e.g., age), and a video risk factor. Here, we emphasize the special importance of clinical factors that have played a traditional role in diagnosis as opposed to a video risk factor that does not have a clear and well understood clinical interpretation within the context of a risk model. Furthermore, to assess the effects of the different modalities, we construct models based on three different sets of variables. First, we consider single modality models based on: (i) CD only (ii) EVM only, and (iii) an EV from parasternal long axis view, which does not include any other measurements. Second, we consider a hierarchy of multi-modal models starting from CD with EVM, and then adding the results from video analysis as well.

We consider polynomial transformations applied to each scalar factor separately as given by:

$$P(X_s) = [p_1(x_1), p_2(x_2), \dots, p_r(x_r)]^T$$

and similarly for $P(X_v)$, where $p(x_i) = \sigma(v_0 + v_1x_i + v_2x_i^2 + v_3x_i^3 + \dots)$ defined in $p : [-1, 1] \rightarrow [0, 1]$.

We use a weighted sum of the contributions from each polynomial based on:

$$W_s^T P(X_s) = w_1 p_1(x_1) + w_2 p_2(x_2) + \dots + w_r p_r(x_r) \quad (4.1)$$

that satisfies

$$W_s = [w_1, w_2, \dots, w_r], \forall w_i \geq 0, \quad (4.2)$$

and similarly for $W_v^T X_v$. Here, we require positive weights to eliminate any model ambiguities in the direction of effect in $p_i(x_i)$ since $w_i p_i(-x_i) = w_i - w_i p_i(x_i)$.

For binary variables, we simplify $W_b^T P_b(X_b) = W_b^T X_b$ and remove the non-negative constraint for W_b .

Chapter 4. Multimodal Interpretable Risk Prediction

For each modality, we consider a sigmoid for modeling the risk likelihood. We thus have that the CD scalar and EVM models are given by:

$$m_s(X_s) = \sigma(W_s^T P(X_s) + b_s) \tag{4.3}$$

$$m_v(X_v) = \sigma(W_v^T P(X_v) + b_v) \tag{4.4}$$

where b_s, b_v represent bias terms, and $\sigma(\cdot)$ represents the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$. We use a binary cross-entropy cost function to train the different models and learn the polynomial weights, coefficient weights, and bias terms.

For the hierarchical, multi-modality models, we consider the original scalar model ($m_s(X_s)$), a second model that also considers EVM: $m_{sv}(\cdot)$, and the full multi-modality model: $m_{sV}(\cdot)$. To simplify the notation, we use the same weight variables to define $m_{sv}(\cdot)$ and $m_{sV}(\cdot)$ as given by:

$$m_{sv}(X_s, X_v) = \sigma(W_s^T P(X_s) + W_v^T P(X_v) + b_{sv}) \tag{4.5}$$

$$m_{sV}(X_s, X_v, V) = \sigma(W_s^T P(X_s) + W_v^T P(X_v) + w_V^T V + b_{sV}) \tag{4.6}$$

where b_{sv}, b_{sV} represent bias terms, and the weights W_s, W_v, w_V will need to be learned for the new models.

4.4.1 Feature Importance

We provide interpretation of the proposed methodology based on separability that allows comparisons among the multi-modal input features, whether scalar, binary or video. We begin with interpreting the contributions of scalar features. We then proceed with looking at the relative importance of the different features within the different models.

Each scalar feature contributes to the overall mortality risk through its corre-

Chapter 4. Multimodal Interpretable Risk Prediction

sponding coefficient weight that is then input to a logistic regression layer,

$$m_s(X_s) = \frac{1}{1 + \exp(-W_s^T(P(X_s) + b_s))} \quad (4.7)$$

that gives a risk score. Since we are using the logistic regression $\sigma(\cdot)$, from

$$\sigma(\text{logit}(p_{\text{mort}})) = p_{\text{mort}},$$

where p_{mort} represents the event probability, we have:

$$\begin{aligned} W_s^T P(X_s) + b_s &= \text{logit}(p_{\text{mort}}) \\ &= \log\left(\frac{p_{\text{mort}}}{1 - p_{\text{mort}}}\right) \end{aligned} \quad (4.8)$$

where $p_{\text{mort}}/(1 - p_{\text{mort}})$ represents the odds ratio for the event. We say that the product $W_s^T P(X_s) + b_s$ represents the log-odds of a mortality event [107]. To understand the risk contribution for the i -th feature, we exponentiate both sides of eq. (4.8) to eventually derive:

$$\text{Odds-ratio} = C \cdot \exp(w_i p_i(x_i)) \quad (4.9)$$

where C represents contributions from the rest of the features. From eq. (4.9), we can see how the weight magnitude can be used to quantify specific feature contributions to the odds ratio. We will refer to eq. (4.9) in the results.

We rank the importance of each feature by simply ranking the corresponding weights: $|w|_{(1)} \geq |w|_{(2)} \geq \dots \geq 0$. Here, it is important to note that eq. (4.9) describes the contribution of each factor over the entire range of possible values, while being invariant to the input scales. The scale invariance is given by the $\sigma(\cdot)$ applied to each polynomial transformation.

Large-scale changes can be described by looking at the change from $\exp(w_i p_i(x_i))$ to $\exp(w_i p_i(x_i + \Delta x_i))$ where Δx_i is used to describe a large change in x_i .

4.4.2 Direction of effect

Since binary variables pass directly to an unconstrained coefficient, the sign of the coefficient would indicate the direction of effect. Then, as described in eq. (4.9), a change from 0 to 1 in the binary input results in an odds ratio change of $\exp(w_i)$. If $w_i > 0$ then $\exp(w_i) > 1$ shows an increase in risk odds. Conversely, if $w_i < 0$ then $0 < \exp(w_i) < 1$ shows a decrease in risk odds.

For continuous variables, $p_i(x_i)$ describes the relation between input, x_i , and its contribution to risk relative to the coefficient w_i . Similarly to a binary variable, the range of $p(\cdot)$ is $[0, 1]$, thus the constrained $w_i > 0$ will determine the increase in risk as $p(\cdot)$ increases.

The lowest risk for each variable can be then determined by

$$\begin{aligned} x_i^{(0)} = \arg \min_{x_i} p_i(x_i) \\ \text{subject to } -1 \leq x_i \leq 1 \end{aligned} \tag{4.10}$$

Equivalently, the maximum risk is defined as

$$\begin{aligned} x_i^{(1)} = \arg \max_{x_i} p_i(x_i) \\ \text{subject to } -1 \leq x_i \leq 1 \end{aligned} \tag{4.11}$$

The polynomial defined in the domain of x , $[-1, 1]$, will have a range of $[0, 1]$, same as a binary variable. Thus, the coefficient now indicates a unit increase in $p(\cdot)$ instead of a unit increase in x as a standard Logistic regression would show.

4.4.3 Multimodal Assessment

To understand the contributions from the EVMs, we rely on the joint interpretation of our hierarchical models: $m_s(\cdot)$, $m_{sv}(\cdot)$, $m_{sV}(\cdot)$. As long as the different models con-

Chapter 4. Multimodal Interpretable Risk Prediction

tribute information associated with the label, we expect the performance to follow the hierarchy with $m_{sV}(\cdot)$ giving the best results, followed by $m_{sv}(\cdot)$, and then either of $m_s(\cdot)$, $m_v(\cdot)$, or $m_V(\cdot)$. The relative improvement in performance can be attributed to the added information in each model. Thus, the performance improvement of $m_{sv}(\cdot)$ over $m_s(\cdot)$ is directly attributed to the inclusion of EVM. Similarly, a performance improvement of $m_{sV}(\cdot)$ over $m_{sv}(\cdot)$ implies that the video analysis system is extracting important features that are currently not fully described by the EVM included in $m_{sv}(\cdot)$. Here, we note that a substantial improvement of $m_{sV}(\cdot)$ over $m_{sv}(\cdot)$ may imply that the current clinical EVM are incomplete. On the other hand, the lack of a substantial improvement may be due to the fact that the video processing system was unable to provide new information that could surpass the standard EVM that we are already making, on the context of one-year mortality prediction. In the case of redundant information, the coefficients of redundant inputs could lean to the super set variable. Moreover, potential lack of improvement in this scenario could also be due to the fact that we are only including one video out of an average of 20-40 videos acquired per clinical echocardiography due to computational limitations.

Beyond performance improvements, we look at changes in weights and weight rankings to assess the importance of each feature from each modality. Performance changes reflect contributions from each modality as a whole, while weight rankings reflect the relative importance of each feature against all others. The presence of high-ranking features from all modalities implies that each modality is making a significant contribution. Further, the relative rankings also matter. For example, the presence of video that ranks higher than echocardiography measurements implies that the video score contains information already given by EVM but within a simpler, single number. Similarly, weight rank changes between models can offer strong clues about the inter-relationships between clinical factors and different modalities.

4.4.4 Training, Validation, and Testing

To estimate the performance of the different models, we performed 5 independent runs. For each run, the dataset was broken into a training set, a validation set, and the test set. We train over the training and validation sets. We report the results over the 5 test sets.

For each run, we used 80% of the dataset for training and validation and the remaining 20% for testing. Within the 80% reserved for training and validation, we used 10% (8% of the original dataset) for validation and the rest for fitting. The training and test sets had the same prevalence of dead vs alive, see Table 4.2. For the validation set, we used a balanced proportion of 50% for each class of Table 4.2.

We normalize each feature by mapping its minimum value to -1 and the maximum value to +1 on the training set using:

$$x_{i,\text{nor}} = 2 \cdot \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} - 1 \quad (4.12)$$

Then, we apply eq. (4.12) to the validation and test sets with the minimum and maximum values found on the training set.

To account for sample imbalance, we weigh the error contributions based on the number of samples in each class as in:

$$i^{\text{th}} \text{ Class Error Weight} = \frac{\text{Total number of samples}}{2(\text{Number of samples in class } i)}$$

Thus, we weight the error of the < 1 year class by 3.14, and the ≥ 1 year by 0.59.

Over the training set, we use the RMSProp optimization method [79] to minimize the binary cross-entropy loss. We trained for a maximum of 10,000 epochs with early stopping if there was no reduction in the validation set loss over 100 consecutive epochs. We implemented all the experiments in Tensorflow (version 1.13). All training was performed on an NVIDIA DGX-1 platform with 8 V100 32GB GPUs.

Table 4.2: Demographics table of 31,278 EHR samples.

	Survival	
	< 1 year	≥ 1 year
Count	4,977	26,301
Male (%)	50	56
Smoker (%)	65	59
Age (years)	73 ± 13	63 ± 16
Heart Rate (bpm)	81 ± 16	73 ± 14
EF (%)	51 ± 14	55 ± 10
LDL (mg/dL)	85 ± 32	93 ± 32
Diastolic Press. (mm[Hg])	66 ± 14	72 ± 13
Systolic Press. (mm[Hg])	124 ± 23	131 ± 21

4.5 Dataset

This retrospective study was approved by the Geisinger Institutional Review Board and performed with a waiver of consent.

4.5.1 Electronic Health Records

At the time of the study, Geisinger’s echocardiography database contained 594,862 studies from 272,280 unique patients performed over 19 years (February 1998 to September 2018). Each study included patient identifiers, date, and a findings report. Geisinger’s Phenomics Initiative database has modeled these study data into tabular format with human-derived echocardiography measurements, where each row represents a sample and columns the measurement type. Multiple patient encounters were treated independently.

We retrieved the closest (before or after) fasting LDL, HDL, blood pressure, heart rate, and weight measurements that were not taken at the time of the Echocardiography study within a six-month window. When no measurement was available in that

Chapter 4. Multimodal Interpretable Risk Prediction

time window, we set the variable as missing. We included International Classification of Diseases codes (tenth revision) for diseases of the circulatory system, chronic kidney disease, dyslipidemia, and congenital heart defects, were formatted as indicator variables that indicated positive diagnosis at the time of echocardiography.

All measurements were cleaned from physiologically out of limit values, which may be caused by input errors. In cases where no limits could be defined for a measurement, we removed extreme outliers that met two rules: 1) Value beyond the mean plus or minus three standard deviations and 2) Value below the 25th percentile minus 3 interquartile ranges or above the 75th percentile plus 3 interquartile ranges. The outlier values were set as missing.

To support our models, we also needed to deal with missing values. We filled in the missing data with two steps. First, we conducted a time interpolation to fill in missing measurements using all available studies of an individual patient, i.e., missing values in between echocardiography sessions were linearly interpolated if complete values were found in adjacent echocardiography studies acquired before and after the study with a missing value. Then, we kept 115 out of the 480 measurements because they were the most commonly measured with less than 90% missing values. This enabled us to conduct a robust Multiple Imputation by Chained Equations (MICE) [31].

After imputation of the continuous measurements, we imputed the missing diastolic function (which is either normal, abnormal or graded from 1 to 3 in severity) assessment by training a logistic regression classifier (One-vs-All) using 278,160 studies where diastolic function was known. We coded the reported diastolic function in an ordinal fashion with -1 for normal, 0 for dysfunction (but no grade reported), and 1, 2 and 3 for diastolic dysfunction grades I, II, and III, respectively. We calculated the patient's age and survival time from the date of the echocardiogram. The patient status (dead/alive) was based on the last known living encounter or confirmed

death date, which is regularly checked against national death index databases in our system.

While imputation does create artificial measurements, its effect has been found minimal both by an independent study [108] and with a Random Forest classifier in [24].

4.5.2 Echocardiography videos

An Echocardiography study consists of typically 20–40 ultrasound videos containing multiple views of the heart and vessels with different orientations. We refer to chapter 3 for details in the video extraction and view labeling procedure.

From the echocardiography exams, we kept only the parasternal long axis view since 1) this view is regarded as the most useful view by cardiologists due to being able to capture a large part of the heart’s anatomy in a single view, 2) in chapter 3, this view gave the best performance for predicting the risk of one-year mortality, and 3) including additional videos remained computationally challenging because of the ratio of available samples vs number of parameters to train.

We linearly interpolated all raw videos to a time resolution of 30 frames per second. We then cropped/padded each video to 60 frames (2 seconds).

4.5.3 Clinical and Video Data Merge

We linked the clinical data (CD) and imaging data, and discarded any unlinked data. We gathered 31,278 videos from 26,793 patients. We limited our video sample size from the 594,862 studies available due to storage and time limitations. The CD variables were age, smoking status (ever smoked), sex, diastolic pressure, systolic

Chapter 4. Multimodal Interpretable Risk Prediction

pressure, heart rate, height, weight, low-density lipoprotein (LDL), and high-density lipoprotein (HDL). Finally, we removed patients with less than 1 year of follow-up and randomly selected a single study per patient. Refer to Table 4.2 for a summary of the merged dataset.

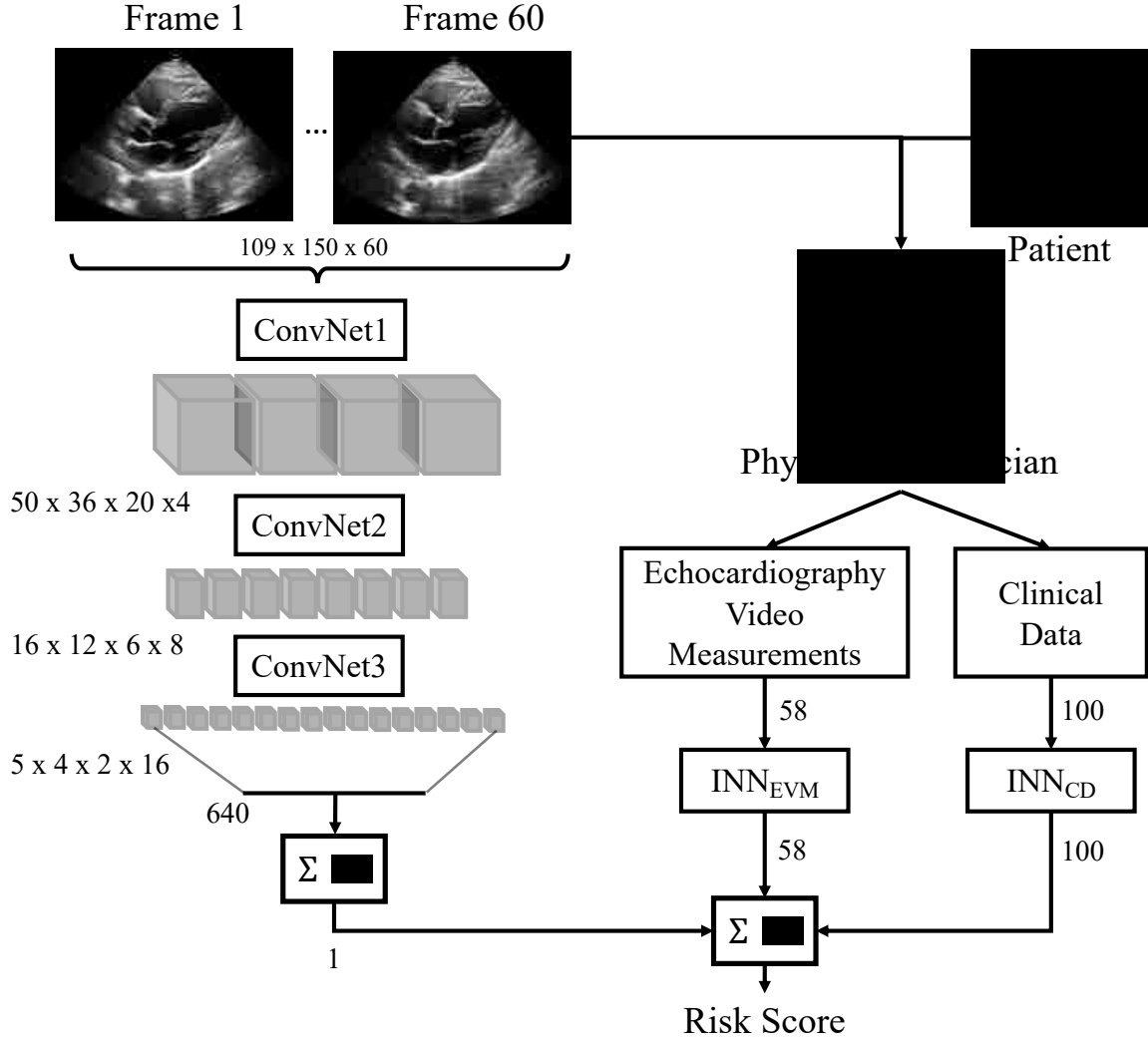


Figure 4.2: Data flow from input to risk score calculation for the proposed multi-modal system. The input is based on an Echocardiography exam and other clinical information (height, weight, etc.). The physician/technician then reads and generates measurements from the video and clinical data from the patient's exam. The output of the 3D CNN video analysis system is connected directly to the final layer of the model. The measurements and clinical data are transformed with the proposed Interpretable Neural Network (INN) which learns 3rd order polynomial transformations that can then contribute to the final risk score. ConvNet[1,2,3] are described in Table 3.4.

Table 4.3: Top five features with their corresponding coefficients for different models. For each coefficient, we provide 95% confidence intervals in between square brackets. Confidence intervals were computed using five folds.

Rank	CD			EVM			CD+EVM			CD+EVM+EV		
	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
1	Age	3.6 [3.1, 4.2]	Ejection Fraction	1.9 [1.7, 2.2]	Age	3.2 [2.9, 3.5]	EV	2.4 [1.9, 2.9]				
2	Heart Rate	2.7 [2.5, 2.9]	Tricuspid Reg MV	1.7 [1.5, 1.8]	Weight	2.0 [1.8, 2.3]	Age	0.6 [0.4, 0.9]				
3	Weight	1.7 [1.5, 1.9]	End Systolic Volume	1.6 [1.3, 2.0]	Heart Rate	1.9 [1.8, 2.0]	Heart Rate	0.3 [0.2, 0.5]				
4	Diastolic Press	1.6 [1.5, 1.8]	LV dim end dias	1.4 [1.1, 1.8]	LVPWd	1.6 [0.6, 2.6]	Tricuspid Reg MV	0.3 [0.1, 0.5]				
5	Systolic Press	1.3 [1.2, 1.5]	AI dec slope	1.2 [1.0, 1.4]	Tricuspid Reg MV	1.5 [1.4, 1.6]	RAP systole	0.3 [0.2, 0.4]				

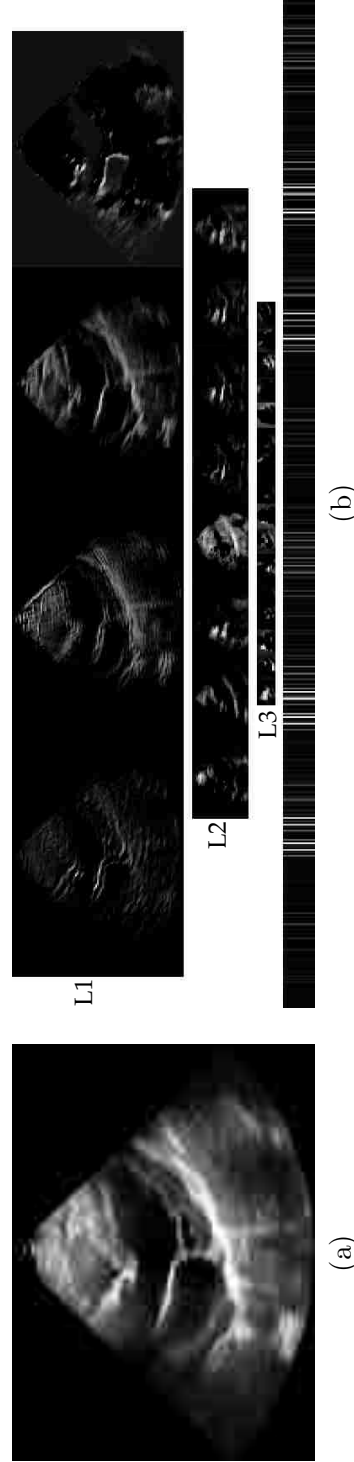


Figure 4.3: Example of low-level features extracted from the parasternal long axis view. In this example, we show (a) a frame of the input video and (b) all outputs from the four feature maps produced by L1 (top), L2 (middle up), L3 (middle down), and the flattened layer enhanced for visualization (spans 10 rows).

4.6 Results and Discussion

We begin with a discussion of the most significant features in section 4.6.1. We then proceed with a discussion of risk models for individual features in section 4.6.2.

We provide an example of our proposed interpretable neural network in section 4.6.3 and a comparison of the different models in section 4.6.4.

4.6.1 Significant features

We summarize the results for the most significant features for the different models in Table 4.3. As expected for a survival model, age dominates all other features in the basic CD model. Age still dominates even after considering measurements derived from the videos (CD+EVM), and it is the second most important feature after EV in the full model. Heart rate, weight, diastolic pressure, and systolic pressure complete the top 5 clinical factors that produced the highest prediction weights. These clinical features are well known and strongly support the interpretability of our results.

From the EVM (without analyzing the video), the top 4 most important variables (Ejection Fraction, Tricuspid Reg MV, End Systolic Volume, and Left Ventricular dimension at end-diastole) are, statistically, equally important. When combining EVM with clinical data (CD + EVM), Tricuspid Reg MV was the only remaining of the top 5 EVM variables. Here, in terms of contributions to the mortality risk, we note that the Tricuspid Reg MV measures the maximum velocity of blood flowing backwards from the right ventricle into the right atrium, which is an indirect measure of pulmonary artery systolic pressure and thus a marker of pulmonary hypertension. Pulmonary hypertension is highly correlated to mortality, as previously discussed in detail in [24], and thus this further supports clinical interpretability of the model.

For the combined model, the video analysis system was the most significant fea-

ture. Unfortunately, it is difficult to provide a clinical interpretation of exactly what is being measured by the video analysis system. In our analysis of video outputs from the lower levels, we have found that the first layers extract granular features, while the subsequent layers show more focused and sparse maps. To show this, we present sample results from the all features maps generated by L1, L2, L3, and the Flatten layers in 4.3 from one of the test videos. The output images from Fig. 4.3b vary significantly. In the top row of Fig. 4.3b, we can see the extraction edges, blurred versions of the input image, and an edge that highlights part of the trajectory of the mitral valve. In the second row, we see several maps highlighting the right ventricle walls. The third row shows more simple and agglomerated bright regions, possibly highlighting key anatomical regions. Finally, the last row shows the static vector that summarizes the entire video. All feature maps were normalized in intensity independently.

4.6.2 Risk model assessment for individual features

We present risk models for the most relevant clinical features, see Table 4.3, in the single modality models for CD and EVM in Fig. 4.4.

We begin with the age factor as a predictor in the CD+EVM model (see Fig. 4.4a). It is clear that mortality increases with age as evidenced by the histogram differences between the two populations (survivors versus non-survivors). In fact, with a weight coefficient of 3.2 (see Table 4.3), we have a 24-fold increase in the odds ratio (probability of dying within a year), when going from an age of 18 to 110. The risk function appears to follow a near linear trend from 40 to 80.

Increased heart rates lead to significant risk increases as shown in Fig. 4.4b. Since these measurements are taken from patients at rest, a low resting heart rate may indicate a physically active and therefore generally healthier person, whereas a

Chapter 4. Multimodal Interpretable Risk Prediction

high rate may be a marker of arrhythmias and/or heart failure.

Extreme low weight gave the highest risk in Fig. 4.4c. From low to average weight, we observe that the risk also drops sharply. The risk drops to the lowest value for patients with weight higher than 250 kilograms. The trend from average to high weight, while appearing to be counter-intuitive, is compatible with the “obesity paradox” noted in multiple prior studies (see [109]). An additional possible explanation is that low weight is a high risk factor for short term (<1 year) mortality and high weight may have a higher association with longer term mortality.

We have decreasing risk trends for larger values of systolic and diastolic pressure (see Figs. 4.4d and 4.4e). Though lower blood pressure being associated with higher risk is counterintuitive, two explanations are plausible. First, a high blood pressure does not lead to 1-year mortality but rather leads to long-term cumulative effects such as renal and heart failure that result in longer-term increased mortality. Second, low blood pressure may be a marker of cardiac decompensation. Full understanding of this trend will require further study and also accounting for many medications known to affect blood pressure.

A mixed trend is observed for the left ventricular ejection fraction (EF) in Fig. 4.4f. It is important to recognize that the majority of the patients fall within the non-linear trend region. It would be a big mistake to suggest a linear trend for the entire ejection fraction region. Here, we note that the EF is the percentage of blood that leaves the heart chambers during contraction. From a minimum risk at 65%, the odds ratio indicates a two-fold increase in risk at an EF of 10% or lower, and a 56% risk increase for an EF of 85% or higher. For low risk cases, the EF risk function agrees with standard clinical interpretation and the current American Heart Association guidelines (reviewed as of May 31, 2017) for a normal EF, which is between 50% and 70%. Increased risk with high EF may be a marker of a hyperdynamic heart failure with preserved ejection fraction or additional pathologic factors known to elevate EF

Chapter 4. Multimodal Interpretable Risk Prediction

such as mitral regurgitation or concentric hypertrophic remodeling (either genetic or acquired secondary to hypertension).

We have a strong, positive trend for increases in the Tricuspid Reg MV (see Fig. 4.4g). Based on our prior discussion on the Tricuspid Reg MV, this trend is clearly to be expected and compatible with pulmonary hypertension being strongly linked to mortality.

From Fig. 4.4h, we can see that higher values of the aortic insufficiency deceleration slope demonstrate the relationship between severity of aortic valve regurgitation/insufficiency and mortality [110]. On the other hand, from Fig. 4.4i, we see a counter-intuitive trend for the left ventricular internal dimension that suggests a lower value is associated with a higher risk of death, which is opposite of what is expected [111]. However, the histograms in Fig. 4.4i show little difference between the surviving and non-surviving populations. Hence, it is not a surprise that this feature was not found to be significant and the trend likely is spurious. From Fig. 4.4j, we observe that the left ventricular end systolic volume, which is a better marker of ventricular size than the LV internal dimension described above, follows the expected trend of worsening mortality for higher values.

Table 4.4: Normalization parameters for eq. (4.13). The minimum and maximum values are used for normalizing the input so that it varies between -1 and +1. The transformation equation is described in eq. (4.12).

	Units	Min	Max
Age	Years	18.0	106.4
Heart Rate (HR)	Bpm	6.0	245.0
Weight (Wt)	Kg	30.4	307.5
Diastolic Pressure (DP)	mmHg	7.0	178.0
Systolic Pressure (SP)	mmHg	4.0	261.0

4.6.3 A fully interpretable Neural Network based on the top-5 clinical data features

To further demonstrate the interpretability of the proposed approach, we provide the full risk assessment model for the top-5 clinical features. From Table 4.3, recall that the top-5 clinical factors are: age, weight, heart rate, diastolic pressure, and systolic pressure. Here, we note that a patient with a weight scale and a blood pressure monitor can actually monitor all of these factors continuously at home.

To compute the risk, we begin by standardizing each factor in the range of -1 to +1 using equation (4.12). For completeness, we provide the min and max values in Table 4.4. We then compute the risk using:

$$\begin{aligned}
 \text{Risk} = & \sigma(4.1 \cdot \sigma(-0.3 + 1.5 \cdot \text{Age} - 1.9 \cdot \text{Age}^2 + 4.7 \cdot \text{Age}^3) + \\
 & 2.5 \cdot \sigma(1.8 + 3.7 \cdot \text{HR} + 3.5 \cdot \text{HR}^2 + 3.7 \cdot \text{HR}^3) + \\
 & 1.3 \cdot \sigma(-2 + 1.8 \cdot \text{Wt} - 1.5 \cdot \text{Wt}^2 + 3.2 \cdot \text{Wt}^3) + \\
 & 1.5 \cdot \sigma(-3.2 + 4.2 \cdot \text{DP} - 5.3 \cdot \text{DP}^2 - 5 \cdot \text{DP}^3) + \\
 & 1.0 \cdot \sigma(-2 + 8.6 \cdot \text{SP} - 12.4 \cdot \text{SP}^2 - 6.1 \cdot \text{SP}^3) + 1)
 \end{aligned} \tag{4.13}$$

where Age , HR , Wt , DP and SP refer to the normalized versions of age, heart-rate, weight, diastolic pressure, and systolic pressure respectively, and $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function.

While slightly different than this reduced model, we refer to Fig. 4.4 for the plots of how each polynomial term affects the full model. Again, we emphasize the unique ability of the proposed approach to capture the global effects of each factor while the full model equation of (4.13) captures the combination of the top-5 features put together.

From equation (4.13), we can see that the non-linear effects are very significant. The non-linear coefficients for the second and third order terms are significantly

higher than zero and cannot be captured by a linear regression model. On the other hand, it is also important to note that due to the fact that we are constraining each factor to $-1 \leq x \leq 1$, the second, third, and higher polynomial terms are dominated by the linear term: $x \leq x^2 \leq x^3 \leq \dots$. Thus, our approach is a generalization of linear regression where we allow linear dependencies to dominate. As discussed earlier, for some factors (e.g., ejection fraction in Fig. 4.4f), a linear model would be highly inappropriate. The simple model represented by equation (4.13) achieves an AUC area of 0.76 compared to the optimal value of 0.83.

4.6.4 Model results

In this section, we provide comparisons across different modalities and different classifiers. We also present our results for multi-modality classification.

To demonstrate the performance of the proposed interpretable network, refer to Fig. 4.5. In Fig. 4.5, we present the AUC as a function of the number of input features for all classifiers. For both CD and EVM, we can see that the Logistic Regression classifier gave the worst performance. Furthermore, the proposed interpretable neural network (INN) approach closely follows the results of powerful non-linear and non-interpretable classifiers (Random Forests and XGBoost).

For classifiers based on all features, a summary of the results is given in Table 4.5. From the results, it is clear that the interpretable neural network performs significantly better than logistic regression while approximating the results of Random Forests and XGBoost. In fact, we have found no statistical difference between the proposed interpretable neural networks and XGBoost or Random Forests.

In terms of single-modality classifiers, the proposed low-parameter dyadic 3D CNN outperformed all other classifiers for the human crafted EVM, which is derived from multiple other videos besides the parasternal long axis view (see Table 4.5).

Chapter 4. Multimodal Interpretable Risk Prediction

Furthermore, the 3D CNN performed significantly better than the 2D CNN classifier as it is clearly documented in Table 4.6. Overall, the combination of the 3D CNN with the interpretable neural network over CD and EVM gave the best overall results with an average AUC of 0.83.

A slightly different performance for the combination of CD+EVM was reported in [24]. We determined that the source of this difference relates to the specific subset of patients included in this analysis (selected based on the availability of the raw echocardiography videos). While the exact cause of the bias is unknown, we do note that our current population, compared to that of [24], did exhibit several demonstrable differences in features, such as 1) increased prevalence of dead patients within a year (16% vs 12%); 2) larger proportion of patients with mild Tricuspid (33% vs 26%) and Mitral (33% vs 25%) Regurgitation; and 3) larger percentages of patients with diagnoses of chronic kidney disease (19% vs 13%), hypertension (54% vs 47%) and heart failure (16% vs 13%).

Also, the proposed approach replicates 5 of the top 10 features reported as most important in [24], for the equivalent of the CD+EVM in this study. The top 10 features in [24] are age, tricuspid regurgitation maximum velocity (Tricuspid Reg MV), heart rate, LDL, pulmonary artery acceleration time, systolic pressure and diastolic function. The proposed approach replicates Age, Tricuspid Reg MV, heart rate, systolic and diastolic pressure.

The advantage of the proposed interpretable neural network approach comes from its ability to describe non-linear relationships for different factors and across modalities. As it is clear from the examples in Fig. 4.4, there are strong non-linear relationships between risk and its dominant clinical factors. Furthermore, it is clear that such non-linear relationships cannot be captured using linear regression models and cannot be easily explained by other non-linear models such as XGBoost and Random Forests.

Chapter 4. Multimodal Interpretable Risk Prediction

Table 4.5: Models performances in percent AUC units. For each method and data combination, we present the average AUC and standard deviation based on 5 independent runs. We use the term Interpretable Neural Network (INN) to refer to the proposed method. The CD input does not include EVM or Video features.

Input	INN (proposed)	Logistic Regression	Random Forest	XGBoost
Single Modality				
CD	79.7 (0.6)	79.2 (0.7)	80.5 (0.4)	80.5 (0.7)
EVM	76.2 (1.1)	74.1 (1.5)	76.5 (1.2)	76.8 (1.2)
EV	78.6 (0.7)	–	–	–
Multiple Modalities				
CD+EVM	82.3 (0.6)	81.2 (0.6)	81.3 (0.8)	82.4 (0.7)
CD+EVM+EV	83.0 (0.4)	–	–	–

Table 4.6: Performance of 2D and 3D CNN video models in percent units for single and multi-modality inputs.

Model Input	IMNN + 2D CNN + LSTM	IMNN + 3D CNN
Echo Video (EV)	73.4 (1.9)	78.6 (0.7)
CD+EVM+EV	81.7 (0.8)	83.0 (0.4)

4.7 Conclusion

This chapter introduces interpretable models for risk assessment in clinical scenarios that demand multi-modal data inputs. Through the use of separable, non-linear models, we are able to quantify the contributions of individual clinical factors to the overall risk. The approach allows us to visualize complex non-linear relationships between changes in each factor and other non-linear models. Overall, the proposed interpretable models matched the performance of more complex non-linear methods and thus demonstrate significant potential for expanding the use of neural networks in medicine.

Chapter 4. Multimodal Interpretable Risk Prediction

In future work, we will investigate the different hyper-parameters of our proposed interpretable neural network approach. More specifically, the polynomial degrees for each feature input needs to be verified with nested cross-validation approaches. Unfortunately, given the large size of our dataset, such extensive experimentation has proven to be computationally prohibitive. On the other hand, the fact that the performance approximates non-linear classifiers implies that significantly higher order polynomial methods need not be considered.

The proposed 3D CNN architecture proved to be very effective for processing echocardiography videos. The 3D CNN classifier outperformed human crafted EVMs. For multi-modal risk assessment, the 3D CNN dominated (higher normalized coefficient) than CD and EVM sub-classifiers. Yet, compared to modern classifiers, the 3D CNN uses a relatively low number of trainable parameters.

The code for implementing the proposed methodology is provided in the DISIML package [112] on <http://github.com/alvarouc/disiml>. We used the TabiMISO class for the tabular data experiments and the VideoSISO class for the video branch.

Chapter 4. Multimodal Interpretable Risk Prediction

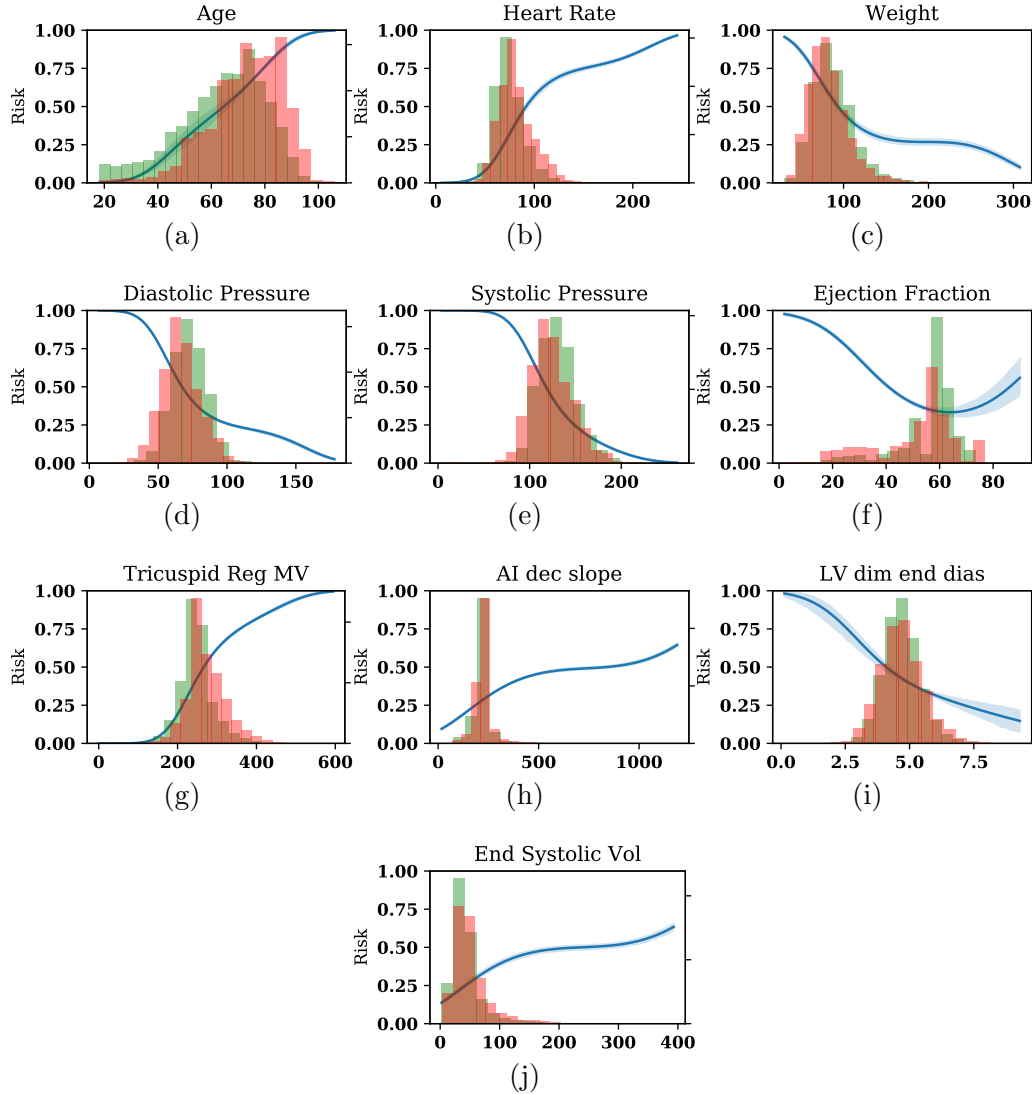


Figure 4.4: Full model risk functions (blue) with normalized histograms of survivors (light green) and non-survivors (red orange). When the two histograms overlap, the histograms appear light brown. Risk function for (a) Age in years, (b) heart rate in beats per minute, (c) weight in kilograms, (d) diastolic and (e) systolic blood pressure in mm Hg, (f) left ventricular ejection fraction in percent, (g) Tricuspid regurgitation maximum velocity in cm/s., (h) aortic insufficiency deceleration slope (AI dec slope) in cm/s^2 , (i) left ventricular internal dimension at end-diastole in cm, and (j) left ventricular end systolic volume in ml. The uncertainty in the risk functions are derived from the 5 results across the 5 runs.

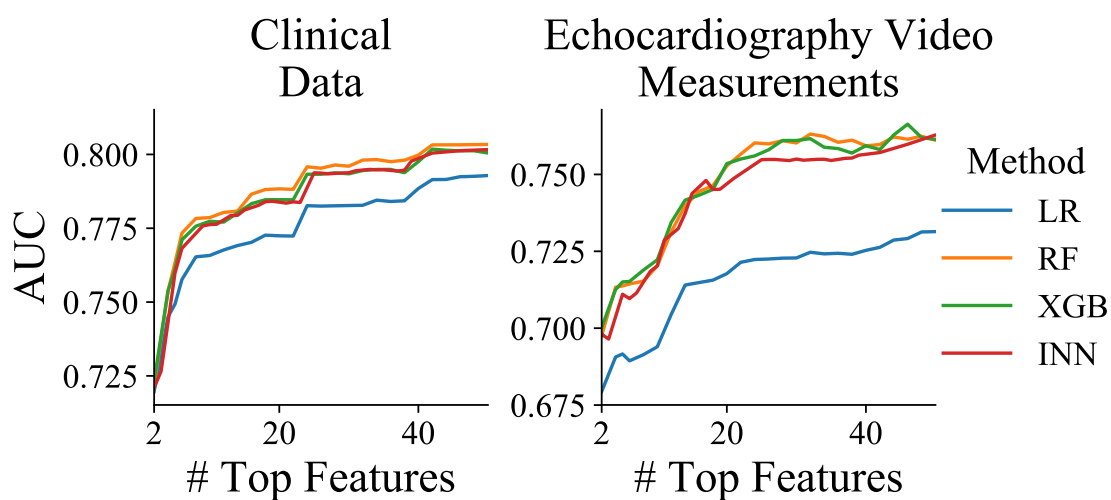


Figure 4.5: AUC performance as a function of the number of the most significant input features for clinical data (left) and echocardiography video measurements (right) for Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), and the proposed method (INN).

Chapter 5

Conclusion and Future Work

We explored and developed methods for EHR data analysis with two approaches, unsupervised and supervised. For the unsupervised model, we did not include the patient survival time information to the model. For the supervised model, we set the labels as indicators of patients that did not survive a year beyond the Echocardiography study.

The unsupervised model labeled patients with similar latent space representations. After a survival analysis of the groups, we found that each group obtained significantly different survival patterns. We then sort the risks of each group, based on the median time of survival and discovered that even a 2-cluster model separated the patients with a larger difference in survival the clinical classification (preserved and reduced Ejection Fraction), which is based in a single Echocardiography measurement.

We then explored the predictive value of Echocardiography videos. We designed four models as combinations of 2D CNNs with LSTMs, 3D CNNs, and GAP layers. We concluded that the best performing model was a 3D CNN model. The best predictive performance can be attained with all views and all Echocardiography

Chapter 5. Conclusion and Future Work

derived measurements (AUC = 0.85). Furthermore, we showed promise of increasing the performance ability even further with learning curves experiments, which doesn't show signs of performance convergence at 30,000 samples.

Finally, we developed a multimodal and interpretable neural network that yielded similar performance as other non-interpretable models for risk prediction from EHR data and is able to incorporate Echocardiography videos. The proposed model learned independent polynomial transformations that described the influence of each variable. The interpretations showed to be clinically consistent and also revealed unexpected trends that require further research.

5.1 Limitations and Future Research

A limitation of the multimodal study was the lack of video samples, which resulted in dropped performance (around 30,000) to match previous results with larger sample sizes (>300,000). The experiments of performance vs sample size, see Fig. 3.10, show evidence that the model could benefit from more samples. As the computational and storage resources become available, I expect to conduct this experiments with sample sizes near 300,000 echocardiography studies.

In the proposed software, I implemented the ability to incorporate selected or all interaction terms. We expected the interaction of variables such as height and weight, and LDL and HDL, to show significance but none was observed. There was no performance gain and it could only match the performance of the model without interaction terms. I also controlled the interaction activation with l1 and l2 regularization terms, but the model solution was in favor of a model with no interaction terms, where the interaction coefficients were all significantly smaller than the regular coefficients. This imposes limitation to the model which would not be able to detect interactions in problems where it is critical. We suggest to manually

Chapter 5. Conclusion and Future Work

identify such interaction with apriori knowledge and handcraft those features.

The software allows for different polynomial order for each input. This creates an immense hyper-parameter search. To simplify the search, I imposed the same order to all inputs and relied on the smallest polynomial order that allowed for monotonical and non-monotonical functions (order 3). For future work, a regularization constrain in the polynomial coefficients may allow a large polynomial order to be reduced based on the influence of each coefficient. This may eliminate the need for a hyper-parameter search.

Currently, a clinical trial in Geisinger is exploring the ability of black-box models to predict one-year mortality on patients with Heart Failure. The model contains “care gaps” as inputs. Care gaps can be understood as treatments, for example a care gap for the flu shot is an indicator variable of whether the patient has taken a flu shot in the current flu season. The proposed model could weight the relative effects of a care gap versus other actionable inputs, such as weight or smoking status, which may allow the exploration of combined treatments.

Lastly, I should seek for an external validation through cross-institution collaboration. This would test the generalizability of our models, especially considering inherent biases of a Geisinger clinical setting, such as the 95% Caucasian population.

5.2 Conclusion

Interpretability is mostly desired for high stakes decision problems. The design of interpretable models allows for auditing and monitoring of the models. I developed an interpretable model with a multi-modal extension that is able to compete in performance with state-of-the-art non-interpretable models. The transparency and interpretability of the model prohibits its secret commercialization. Thus, it will

Chapter 5. Conclusion and Future Work

enable us to gain physician and patient trust to machine learning, which in turn will facilitate the proliferation of open tools to hospitals around the world.

Appendices

Appendix A

Supplementary Tables

Table A.2: Description of all variables extracted from the electronic health records. *MOD = modified ellipsoid, **el = (single plane) ellipsoid, LV = left ventricular, IV = inter-ventricular. ¹⁻¹⁰ Selected EHR variables previously reported as the top 10 predictors of 1-year mortality. *Hot encoded for severity levels 0,1,2,3. Diastolic function coding -1: Normal, 0: abnormal (no grade reported), [1,2,3]: grade I/II/I

	EHR VARIABLE	UNITS	DESCRIPTION
	Demographics		
1	Age ¹	years	At the time of study
2	Sex	0: Female, 1: Male	
3	Smoking status	0: No, 1: Yes	Ever smoked

Appendix A. Supplementary Tables

Vitals			
4	Height	cm	
5	Weight	kg	
6	Heart rate ³	bpm	
7	Diastolic blood pressure ⁶	mm Hg	
8	Systolic blood pressure ⁸	mm Hg	
Laboratory			
9	LDL ⁴	mg/DL	Low-density lipoprotein
10	HDL	mg/DL	High-density lipoprotein
Echocardiography measurement			
11	LVEF ⁵	%	Physician-reported left ventricular ejection fraction
12	AI dec slope	cm/s ²	Aortic insufficiency deceleration slope
13	AI max vel	cm/s	Aortic insufficiency maximum velocity
14	Ao V2 VTI	cm	Velocity-time integral of distal to aortic valve flow
15	Ao V2 max	cm/s	Maximum velocity of distal to aortic valve flow
16	Ao V2 mean	cm/s	Mean velocity of distal to aortic valve flow
17	Ao root diam	cm	Aortic root diameter
18	Asc Aorta	cm	Ascending aortic diameter

Appendix A. Supplementary Tables

19	EDV MOD*- sp2	ml	LV end-diastolic volume: apical 2-chamber
20	EDV MOD*- sp4	ml	LV end-diastolic volume: apical 4-chamber
21	EDV sp2-el**	ml	LV end-diastolic volume: apical 2-chamber
22	EDV sp4-el**	ml	LV end-diastolic volume: apical 4-chamber
23	ESV MOD*-sp2	ml	LV end-systolic volume: apical 2-chamber
24	ESV MOD*-sp4	ml	LV end-systolic volume: apical 4-chamber
25	ESV sp2-el**	ml	LV end-systolic volume: apical 2-chamber
26	ESV sp4-el**	ml	LV end-systolic volume: apical 4-chamber
27	IVSd	cm	IV septum dimension at end-diastole
28	LA dimension	cm	Left atrium dimension
29	LAV MOD*-sp2	ml	Left atrium volume: apical 2-chamber
30	LAV MOD*-sp4	ml	Left atrium volume: apical 4-chamber
31	LV V1 VTI	cm	Velocity-time integral: proximal to the obstruction
32	LV V1 max	cm/s	Maximum LV velocity: proximal to the obstruction

Appendix A. Supplementary Tables

33	LV V1 mean	cm/s	Mean LV velocity proximal to the obstruction
34	LVAd ap2	cm ²	LV area at end-diastole: apical 2-chamber
35	LVAd ap4	cm ²	LV area at end-diastole: apical 4-chamber
36	LVA _s ap2	cm ²	LV area at end-systole: apical 2-chamber
37	LVA _s ap4	cm ²	LV area at end-systole: apical 4-chamber
38	LVIDd	cm	LV internal dimension at end-diastole
39	LVID _s	cm	LV internal dimension at end-systole
40	LVLd ap2	cm	LV long-axis length at end-diastole: apical 2-chamber
41	LVLd ap4	cm	LV long-axis length at end-diastole: apical 4-chamber
42	LVL _s ap2	cm	LV long-axis length at end-systole: apical 2-chamber
43	LVL _s ap4	cm	LV long-axis length at end-systole: apical 4-chamber
44	LVOT area M	cm ²	LV outflow tract area
45	LVOT diam	cm	LV outflow tract diameter
46	LVPWd	cm	LV posterior wall thickness at end-diastole
47	MR max vel	cm/s	Mitral regurgitation maximum velocity

Appendix A. Supplementary Tables

48	MV A point	cm/s	A-point maximum velocity of mitral flow
49	MV E point	cm/s	E-point maximum velocity of mitral flow
50	MV P1/2t max-vel	cm/s	Maximum velocity of mitral valve flow
51	MV dec slope	cm/s ²	Mitral valve deceleration slope
52	MV dec time	s	Mitral valve deceleration time
53	PA V2 max	cm/s	Maximum velocity of distal to pulmonic valve flow
54	PA acc slope ⁹	cm/s ²	Pulmonary artery acceleration slope
55	PA acc time ⁷	s	Pulmonary artery acceleration time
56	Pulm. R-R	s	Pulmonary R-R time interval
57	RAP systole	mm-Hg	Right atrial end-systolic mean pressure
58	RVDd	cm	Right ventricle dimension at end-diastole
59	TR max vel ²	cm/s	Tricuspid regurgitation maximum velocity
60	AVR	0/1*	Aortic valve regurgitation
61	MVR	0/1*	Mitral valve regurgitation
62	TVR	0/1*	Tricuspid valve regurgitation
63	PVR	0/1*	Pulmonary valve regurgitation
64	AVS	0/1*	Aortic valve stenosis

Appendix A. Supplementary Tables

65	MVS	0/1*	Mitral valve stenosis
66	TVS	0/1*	Tricuspid valve stenosis
67	PVS	0/1*	Pulmonary valve stenosis
68	Diastolic function ¹⁰	-1,01,2,3,4	Physician-reported diastolic function
Diagnosis codes			
69-71	I00, I01, I02		Acute rheumatic fever
72-76	I05, I06, I07, I08, I09		Chronic rheumatic heart disease
77-82	I10, I11, I12, I13, I15, I16		Hypertensive diseases
83-88	I20, I21, I22, I23, I24, I25		Ischemic heart diseases
89-91	I26, I27, I28		Pulmonary heart disease and diseases of pulmonary circulation
92	I30		Acute pericarditis
93-106	I31, I32, I33, I34, I35, I36, I37, I38, I39, I43, I44, I45, I49, I51		Other forms of heart disease
107	I40		Acute myocarditis
108	I42		Cardiomyopathy
109	I46		Cardiac arrest
110	I47		Paroxysmal tachycardia
111	I48		Atrial fibrillation
112	I50		Heart failure

Appendix A. Supplementary Tables

113-121	I60, I61, I62, I63, I65, I66, I67, I68, I69	Cerebrovascular diseases
122-131	I70, I71, I72, I73, I74, I75, I76, I77, I78, I79	Diseases of arteries, arterioles and capillaries
131-140	I80, I81, I82, I83, I85, I86, I87, I88, I89	Diseases of veins, lymphatic vessels, and lymph nodes
141	I95	Hypotension
142-144	I96, I97, I99	Other and unspecified disor- ders of the circulatory system
145-149	E08, E09, E10, E11, E13	Diabetes mellitus
150-156	Q20, Q21, Q22, Q23, Q24, Q25, Q26	Congenital heart defect
157	E78	Dyslipidemia
158	N18	Chronic kidney disease

Appendix A. Supplementary Tables

Table A.3: Number of valid samples after setting 600 studies aside for the final test comparison to the 2 cardiologists.

VIEW GROUP/MONTHS	3	6	9	12
Apical 2	19,334	19,328	19,323	19,316
Apical 3	19,392	19,388	19,384	19,376
Apical 4	18,755	18,749	18,745	18,737
Apical 4 Focused to RV	21,192	21,186	21,181	21,173
Apical 5	18,438	18,431	18,426	18,419
Parasternal Long Axis	22,426	22,420	22,415	22,407
Parasternal Long Ascending AORTA	21,700	21,694	21,688	21,681
Parasternal Long RV Inflow	21,544	21,538	21,534	21,528
Parasternal Long Zoom Aor- tic Valve	21,657	21,650	21,645	21,637
Parasternal Short Aortic Valve	21,875	21,870	21,865	21,857
Parasternal Short Pulmonic Valve and Pulmonary Artery	21,614	21,609	21,605	21,596
Parasternal Short Tricuspid Valve	13,385	13,379	13,375	13,370
Short Axis Base	21,541	21,535	21,530	21,523
Subcostal 4 Chamber	20,768	20,763	20,758	20,751
Subcostal Hepatic Vein	11,033	11,029	11,024	11,020
Subcostal Inter-Atrial Sep- tum	19,402	19,399	19,394	19,387

Appendix A. Supplementary Tables

Subcostal IVC with Respiration	20,510	20,505	20,499	20,492
Subcostal RV	20,263	20,259	20,254	20,247
Suprasternal Notch	18,382	18,378	18,372	18,365
Short Axis Mid Papillary	21,801	21,796	21,791	21,783
Short Axis Apex	21,870	21,864	21,859	21,851

Appendix A. Supplementary Tables

Table A.1: View labels found in DICOM tags for the corresponding view type. The view tag in bold indicates the abbreviation used for the view type.

VIEW TYPE	VIEW TAGS
Apical 2	a2, ap2 2d, a2 2d, a2 lavol, la 2ch
Apical 3	a long, ap3 2d, a3 2d
Apical 4	ap4, ap4 2d, a4 2d, a4 zoom, a4 lavol, la ap4 ch
Apical 4 focused to rv	rv focus, rvfocus
Apical 5	a5, ap5 2d, a5 2d
Parasternal long axis	pl deep, psl deep
Parasternal long ascending aorta	pl ascao, asc ao, pl asc ao
Parasternal long mitral valve	pla mv
Parasternal long pulmonic valve	pl pv, pv lax
Parasternal long rv inflow	pl rvif, rv inf, rvif 2d
Parasternal long zoom aortic valve	pl av ao, av zoom
Parasternal short aortic valve	ps av, psavzoom, psax av
Parasternal short pulmonic valve and pulmonary artery	ps pv pa, ps pv, psax pv
Parasternal short tricuspid valve	ps tv, ps tv 2d, psax tv
Short axis apex	sax apex
Short axis base	lv base
Short axis mid papillary	sax mid, sax
Subcostal 4 chamber	sbc 4 ch, sbc 4, sbc 4ch
Subcostal hepatic vein	ivc hv, sbc hv
Subcostal inter-atrial septum	ias, sbc ias, ias 2d
Subcostal ivc with respiration	ivc resp, sbc ivc, ivc insp, ivc sniff, ivcsniff, sniff
Subcostal rv	sbc rv
Suprasternal notch	ssn, ssn sax
Parasternal long lax	lax
Short axis mid papillary	lv mid
Short axis apex	lv apex
Apical 3 zoom	ap3
Apical 2 zoom	ap2
Short axis base	sax base

References

- [1] C. A. McCarty, R. L. Chisholm, C. G. Chute, I. J. Kullo, G. P. Jarvik, E. B. Larson, R. Li, D. R. Masys, M. D. Ritchie, D. M. Roden *et al.*, “The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies,” *BMC medical genomics*, vol. 4, no. 1, p. 13, 2011.
- [2] I. Sutskever, G. E. Hinton, and A. Krizhevsky, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [5] K. D. Mandl and I. S. Kohane, “Escaping the ehr trapthe future of health it,” *New England Journal of Medicine*, vol. 366, no. 24, pp. 2240–2242, 2012.
- [6] S. J. Aronson and H. L. Rehm, “Building the foundation for genomics in precision medicine,” *Nature*, vol. 526, no. 7573, p. 336, 2015.
- [7] S. Kenchaiah, J. C. Evans, D. Levy, P. W. Wilson, E. J. Benjamin, M. G. Larson, W. B. Kannel, and R. S. Vasan, “Obesity and the risk of heart failure,” *New England Journal of Medicine*, vol. 347, no. 5, pp. 305–313, 2002.
- [8] W. C. Levy, D. Mozaffarian, D. T. Linker, S. Sutradhar, S. Anker, A. Cropp, I. Anand, A. Maggioni, P. Burton, M. Sullivan *et al.*, “The seattle heart failure model,” *Circulation*, vol. 113, no. 11, pp. 1424–1433, 2006.

References

- [9] A. T. Yan, P. Jong, R. T. Yan, M. Tan, D. Fitchett, C.-M. Chow, M. T. Roe, K. S. Pieper, A. Langer, S. G. Goodman *et al.*, “Clinical trial-derived risk model may not generalize to real-world patients with acute coronary syndrome,” *American heart journal*, vol. 148, no. 6, pp. 1020–1027, 2004.
- [10] C. G. Victora, J.-P. Habicht, and J. Bryce, “Evidence-based public health: moving beyond randomized trials,” *American journal of public health*, vol. 94, no. 3, pp. 400–405, 2004.
- [11] S. Van Poucke, M. Thomeer, J. Heath, and M. Vukicevic, “Are randomized controlled trials the (g) old standard? from clinical intelligence to prescriptive analytics,” *Journal of medical Internet research*, vol. 18, no. 7, p. e185, 2016.
- [12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [14] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.
- [15] M. C. Mukkamala and M. Hein, “Variants of rmsprop and adagrad with logarithmic regret bounds,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2545–2553.
- [16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.
- [17] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.
- [18] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson,

References

- S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.
- [19] L. Yu, Y. Guo, Y. Wang, J. Yu, and P. Chen, “Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1886–1895, 2017.
- [20] J. Zhang, S. Gajjala, P. Agrawal, G. H. Tison, L. A. Hallock, L. Beussink-Nelson, M. H. Lassen, E. Fan, M. A. Aras, C. Jordan *et al.*, “Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy,” *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.
- [21] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng, “Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 507–514.
- [22] H. Chen, L. Wu, Q. Dou, J. Qin, S. Li, J.-Z. Cheng, D. Ni, and P.-A. Heng, “Ultrasound standard plane detection using a composite neural network framework,” *IEEE transactions on cybernetics*, vol. 47, no. 6, pp. 1576–1586, 2017.
- [23] P. W. Wilson, R. B. D’Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, “Prediction of coronary heart disease using risk factor categories,” *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.
- [24] M. D. Samad, A. Ulloa, G. J. Wehner, L. Jing, D. Hartzel, C. W. Good, B. A. Williams, C. M. Haggerty, and B. K. Fornwalt, “Predicting survival from large echocardiography and electronic health record datasets: Optimization with machine learning,” *JACC: Cardiovascular Imaging*, 2018.
- [25] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [26] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Confounding variables can degrade generalization performance of radiological deep learning models,” *arXiv preprint arXiv:1807.00431*, 2018.
- [27] W. Guan, M. Jiang, Y. Gao, H. Li, G. Xu, J. Zheng, R. Chen, and N. Zhong, “Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics,” *The Int. Journal of Tuberculosis and Lung Disease*, vol. 20, no. 3, pp. 402–410, 2016.

References

- [28] S. J. Shah, D. H. Katz, S. Selvaraj, M. A. Burke, C. W. Yancy, M. Gheorghide, R. O. Bonow, C.-C. Huang, and R. C. Deo, “Phenomapping for novel classification of heart failure with preserved ejection fraction,” *Circulation*, vol. 131, no. 3, pp. 269–279, 2015.
- [29] D. H. Katz, R. C. Deo, F. G. Aguilar, S. Selvaraj, E. E. Martinez, L. Beussink-Nelson, K.-Y. A. Kim, J. Peng, M. R. Irvin, H. Tiwari *et al.*, “Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction,” *Journal of Cardiovascular Trans. Research*, pp. 1–10, 2017.
- [30] A. K. Jain, “Data clustering: 50 years beyond k-means,” *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [31] S. Buuren and K. Groothuis-Oudshoorn, “MICE: Multivariate imputation by chained equations in r,” *Journal of statistical software*, vol. 45, no. 3, 2011.
- [32] G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [33] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.
- [34] S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [35] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [36] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in *Advances in neural information processing systems*, 2002, pp. 849–856.
- [37] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
- [38] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

References

- [39] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 881–892, 2002.
- [40] Y. Senbabaoglu, G. Michailidis, and J. Z. Li, “Critical limitations of consensus clustering in class discovery,” *Scientific reports*, vol. 4, 2014.
- [41] D. R. Cox, “Regression models and life-tables,” in *Breakthroughs in statistics*. Springer, 1992, pp. 527–541.
- [42] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Scientific reports*, vol. 6, p. 26094, 2016.
- [43] B. K. Beaulieu-Jones, C. S. Greene *et al.*, “Semi-supervised learning of the electronic health record for phenotype stratification,” *Journal of biomedical informatics*, vol. 64, pp. 168–178, 2016.
- [44] C. Penone, A. D. Davidson, K. T. Shoemaker, M. Di Marco, C. Rondinini, T. M. Brooks, B. E. Young, C. H. Graham, and G. C. Costa, “Imputation of missing data in life-history trait datasets: which approach performs the best?” *Methods in Ecology and Evolution*, vol. 5, no. 9, pp. 961–970, 2014.
- [45] B. K. Beaulieu-Jones, J. W. Snyder, J. H. Moore, S. A. Pendergrass, and C. R. Bauer, “Characterizing and managing missing structured data in electronic health records,” *bioRxiv*, p. 167858, 2017.
- [46] R. B. Devereux, G. De Simone, D. K. Arnett, L. G. Best, E. Boerwinkle, B. V. Howard, D. Kitzman, E. T. Lee, T. H. Mosley, A. Weder *et al.*, “Normal limits in relation to age, body size and gender of two-dimensional echocardiographic aortic root dimensions in persons ≥ 15 years of age,” *The American journal of cardiology*, vol. 110, no. 8, pp. 1189–1194, 2012.
- [47] I. R. White, P. Royston, and A. M. Wood, “Multiple imputation using chained equations: issues and guidance for practice,” *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011.
- [48] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: a critical evaluation,” *BMC medical informatics and decision making*, vol. 16, no. 3, p. 74, 2016.
- [49] J. W. Payne, “Task complexity and contingent processing in decision making: An information search and protocol analysis,” *Organizational behavior and human performance*, vol. 16, no. 2, pp. 366–387, 1976.

References

- [50] G. Quer, E. D. Muse, N. Nikzad, E. J. Topol, and S. R. Steinhubl, “Augmenting diagnostic vision with ai,” *The Lancet*, vol. 390, no. 10091, p. 221, 2017.
- [51] S. Jha and E. J. Topol, “Adapting to artificial intelligence: radiologists and pathologists as information specialists,” *Jama*, vol. 316, no. 22, pp. 2353–2354, 2016.
- [52] E. Kyriacou, A. Constantinides, C. Pattichis, M. Pattichis, and A. Panayides, “eemergency healthcare informatics,” in *Biomedical Signals, Imaging, and Informatics*, 4th ed., J. D. Bronzino and D. Peterson, Eds. CRC Press, 2015, ch. 64.
- [53] M. Neofytou, V. Tanos, I. Constantinou, E. Kyriacou, M. Pattichis, and C. Pattichis, “Computer aided diagnosis in hysteroscopic imaging,” *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, pp. 1–1, 2014. [Online]. Available: <http://ece.unm.edu/ivpcl/Publications/JOURNALS/2014/Computer%20Aided%20Diagnosis%20in%20Hysteroscopic%20Imaging.pdf>
- [54] Z. C. Antoniou, A. S. Panayides, M. Pantzaris, A. G. Constantinides, C. S. Pattichis, and M. S. Pattichis, “Real-time adaptation to time-varying constraints for medical video communications,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1177–1188, 2018.
- [55] A. Panayides, M. Pattichis, C. Loizou, M. Pantziaris, A. Constantinides, and C. Pattichis, “An effective ultrasound video communication system using despeckle filtering and hevc,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 668–676, March 2015. [Online]. Available: <http://ece.unm.edu/ivpcl/Publications/JOURNALS/2015/An%20Effective%20Ultrasound%20Video%20Communication%20System%20Using%20Despeckle%20Filtering%20and%20HEVC.pdf>
- [56] G. Esakki, V. Jatla, and M. S. Pattichis, “Adaptive high efficiency video coding based on camera activity classification,” in *DCC*, 2017, p. 438.
- [57] A. Panayides, C. Loizou, M. Pattichis, E. Kyriacou, C. Shizas, A. Nicolaidis, and C. Pattichis, “Ultrasound video despeckle filtering for high efficiency video coding in m-health systems,” in *CIWSP Workshop (in honor of the 70th birthday of Prof. Constantinides)*, Jan 2013, pp. 1–4. [Online]. Available: http://www.medinfo.cs.ucy.ac.cy/doc/Publications/Conferences/2013/2013-Ultrasound_video_despeckle_filtering_for_high_efficiency_video_coding_in_M-Health_systems.pdf
- [58] A. Panayides, M. Pattichis, C. Pattichis, C. Loizou, and M. Pantziaris, “Wireless ultrasound video transmission for stroke risk assessment: Quality metrics

References

- and system design,” in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2013.
- [59] S. Murillo, V. Murray, C. Loizou, C. Pattichis, M. Pattichis, and S. Barriga, “Motion and deformation analysis of ultrasound videos with applications to classification of carotid artery plaques,” in *SPIE Medical Imaging*, 2012.
- [60] S. Murillo, M. Pattichis, P. Soliz, S. Barriga, C. Loizou, and C. Pattichis, “Global optimization for motion estimation with applications to ultrasound videos of carotid artery plaques,” in *Proc. SPIE Medical Imaging: Computer-Aided Diagnosis*. International Society for Optics and Photonics, 2010, pp. 76 290X–76 290X. [Online]. Available: <http://ece.unm.edu/ivpcl/Publications/2010/Global%20Optimization%20for%20Motion%20Estimation%20with%20Applications%20to%20Ultrasound%20Videos%20of%20Carotid%20Artery%20Plaques.pdf>
- [61] V. Murray, S. Murillo, M. Pattichis, C. Loizou, C. Pattichis, E. Kyriacou, and A. Nicolaides, “An am-fm model for motion estimation in atherosclerotic plaque videos,” in *41st Asilomar Conference on Signals, Systems and Computers*, Nov 2007, pp. 746–750. [Online]. Available: <http://ece.unm.edu/ivpcl/Publications/2007/An%20AM-FM%20model%20for%20Motion%20Estimation%20in%20Atherosclerotic%20Plaque%20Videos.pdf>
- [62] I. Constantinou, M. Pattichis, C. Tziakouri, C. Pattichis, S. Petroudi, and C. Nicosia, “Multiscale am-fm models and instantaneous amplitude evaluation for mammographic density classification.” in *MIUA*, 2014, pp. 271–276.
- [63] C. Loizou, V. Murray, M. Pattichis, M. Pantziaris, and C. Pattichis, “Multiscale amplitude-modulation frequency-modulation (am-fm) analysis of ultrasound images of the intima and media layers of the carotid artery,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 178–188, March 2011. [Online]. Available: http://ivpcl.unm.edu/bibtex_php/Journals_Pdfs/2011/MultiscaleAmplitude.pdf
- [64] C. Agurto, V. Murray, S. Barriga, S. Murillo, M. Pattichis, H. Davis, S. Russell, M. Abramoff, and P. Soliz, “Multiscale am-fm methods for diabetic retinopathy lesion detection,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 502–512, Feb 2010. [Online]. Available: <http://ece.unm.edu/ivpcl/Publications/JOURNALS/2010/Multiscale%20AM-FM%20Methods%20for%20Diabetic%20Retinopathy%20Lesion%20Detection.pdf>
- [65] S. Ji, W. Xu, M. Yang, and K. Yu, “3d convolutional neural networks for human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

References

- [66] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
- [67] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.
- [68] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.
- [69] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, “Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.
- [70] M. R. Arbabshirani, B. K. Fornwalt, G. J. Mongelluzzo, J. D. Suever, B. D. Geise, A. A. Patel, and G. J. Moore, “Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration,” *npj Digital Medicine*, vol. 1, no. 1, p. 9, 2018.
- [71] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P.-A. Heng, “Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks,” *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.
- [72] A. Madani, J. R. Ong, A. Tibrewal, and M. R. Mofrad, “Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease,” *npj Digital Medicine*, vol. 1, no. 1, p. 59, 2018.
- [73] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, “Improving palliative care with deep learning,” *BMC medical informatics and decision making*, vol. 18, no. 4, p. 122, 2018.
- [74] M. Motwani, D. Dey, D. S. Berman, G. Germano, S. Achenbach, M. H. Al-Mallah, D. Andreini, M. J. Budoff, F. Cademartiri, T. Q. Callister *et al.*, “Machine learning for prediction of all-cause mortality in patients with suspected

References

- coronary artery disease: a 5-year multicentre prospective registry analysis,” *European heart journal*, vol. 38, no. 7, pp. 500–507, 2016.
- [75] M. Hadamitzky, S. Achenbach, M. Al-Mallah, D. Berman, M. Budoff, F. Cademartiri, T. Callister, H.-J. Chang, V. Cheng, K. Chinnaiyan *et al.*, “Optimized prognostic score for coronary computed tomographic angiography: results from the confirm registry (coronary ct angiography evaluation for clinical outcomes: An international multicenter registry),” *Journal of the American College of Cardiology*, vol. 62, no. 5, pp. 468–476, 2013.
- [76] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, “Translating videos to natural language using deep recurrent neural networks,” *arXiv preprint arXiv:1412.4729*, 2014.
- [77] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, “Fast and accurate classification of echocardiograms using deep learning,” *arXiv preprint arXiv:1706.08658*, 2017.
- [78] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in neural information processing systems*, 2014, pp. 568–576.
- [79] Y. Dauphin, H. de Vries, and Y. Bengio, “Equilibrated adaptive learning rates for non-convex optimization,” in *Advances in neural information processing systems*, 2015, pp. 1504–1512.
- [80] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [81] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [82] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [83] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4534–4542.
- [84] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

References

- [85] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova *et al.*, “Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging,” *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.
- [86] J. Chesebro, G. Knatterud, R. Roberts, J. Borer, L. Cohen, J. Dalen, H. Dodge, C. Francis, D. Hillis, and P. Ludbrook, “Thrombolysis in myocardial infarction (timi) trial, phase i: A comparison between intravenous tissue plasminogen activator and intravenous streptokinase. clinical findings through hospital discharge.” *Circulation*, vol. 76, no. 1, pp. 142–154, 1987.
- [87] K. A. Eagle, M. J. Lim, O. H. Dabbous, K. S. Pieper, R. J. Goldberg, F. Van de Werf, S. G. Goodman, C. B. Granger, P. G. Steg, J. M. Gore *et al.*, “A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry,” *Jama*, vol. 291, no. 22, pp. 2727–2733, 2004.
- [88] R. W. Foley, R. M. Maweni, L. Gorman, K. Murphy, D. J. Lundon, G. Durkan, R. Power, F. O’Brien, K. J. O’malley, D. J. Galvin *et al.*, “European randomised study of screening for prostate cancer (erspc) risk calculators significantly outperform the prostate cancer prevention trial (pcpt) 2.0 in the prediction of prostate cancer: a multi-institutional study,” *BJU international*, vol. 118, no. 5, pp. 706–713, 2016.
- [89] L. A. Allen, D. D. Matlock, S. M. Shetterly, S. Xu, W. C. Levy, L. B. Portalupi, C. K. McIlvennan, J. H. Gurwitz, E. S. Johnson, D. H. Smith *et al.*, “Use of risk models to predict death in the next year among individual ambulatory patients with heart failure,” *JAMA cardiology*, vol. 2, no. 4, pp. 435–441, 2017.
- [90] A. S. Panayides, M. S. Pattichis, S. Leandrou, C. Pitris, A. Constantinidou, and C. Pattichis, “Radiogenomics for precision medicine with a big data analytics perspective,” *IEEE journal of biomedical and health informatics*, vol. early access, 2018. [Online]. Available: <https://ieeexplore.ieee.org/document/8588324>
- [91] C. Bonner, M. A. Fajardo, S. Hui, R. Stubbs, and L. Trevena, “Clinical validity, understandability, and actionability of online cardiovascular disease risk calculators: Systematic review,” *Journal of medical Internet research*, vol. 20, no. 2, 2018.

References

- [92] M. W. Kattan, K. R. Hess, M. B. Amin, Y. Lu, K. G. Moons, J. E. Gershewald, P. A. Gimotty, J. H. Guinney, S. Halabi, A. J. Lazar *et al.*, “American joint committee on cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine,” *CA: a cancer journal for clinicians*, vol. 66, no. 5, pp. 370–374, 2016.
- [93] D. Bertsimas and J. Dunn, “Optimal classification trees,” *Machine Learning*, vol. 106, no. 7, pp. 1039–1082, 2017.
- [94] Z. Cui, W. Chen, Y. He, and Y. Chen, “Optimal action extraction for random forests and boosted trees,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 179–188.
- [95] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, “Interpretable predictions of tree-based ensembles via actionable feature tweaking,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 465–474.
- [96] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, “Multi-scale convolutional neural networks for lung nodule classification,” in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 588–599.
- [97] M. Hermans and B. Schrauwen, “Training and analysing deep recurrent neural networks,” in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 190–198.
- [98] M. Gevrey, I. Dimopoulos, and S. Lek, “Review and comparison of methods to study the contribution of variables in artificial neural network models,” *Ecological modelling*, vol. 160, no. 3, pp. 249–264, 2003.
- [99] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [100] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *arXiv preprint arXiv:1511.00561*, 2015.
- [101] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

References

- [102] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you?: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.
- [103] —, “Anchors: High-precision model-agnostic explanations,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [104] —, “Nothing else matters: Model-agnostic explanations by identifying prediction invariance,” 2016.
- [105] N. Puri, P. Gupta, P. Agarwal, S. Verma, and B. Krishnamurthy, “Magix: Model agnostic globally interpretable explanations,” *arXiv preprint arXiv:1706.07160*, 2017.
- [106] T. J. Hastie, “Generalized additive models,” in *Statistical models in S*. Routledge, 2017, pp. 249–307.
- [107] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer series in statistics New York, 2009.
- [108] A. Farhangfar, L. Kurgan, and J. Dy, “Impact of imputation of missing values on classification error for discrete data,” *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.
- [109] C. J. Lavie, A. De Schutter, P. Parto, E. Jahangir, P. Kokkinos, F. B. Ortega, R. Arena, and R. V. Milani, “Obesity and prevalence of cardiovascular diseases and prognosis—the obesity paradox updated,” *Progress in cardiovascular diseases*, vol. 58, no. 5, pp. 537–547, 2016.
- [110] W. A. Zoghbi, M. Enriquez-Sarano, E. Foster, P. A. Grayburn, C. D. Kraft, R. A. Levine, P. Nihoyannopoulos, C. M. Otto, M. A. Quinones, H. Rakowski *et al.*, “Recommendations for evaluation of the severity of native valvular regurgitation with two-dimensional and doppler echocardiography,” *Journal of the American Society of Echocardiography*, vol. 16, no. 7, pp. 777–802, 2003.
- [111] R. M. Lang, M. Bierig, R. B. Devereux, F. A. Flachskampf, E. Foster, P. A. Pellikka, M. H. Picard, M. J. Roman, J. Seward, J. Shanewise *et al.*, “Recommendations for chamber quantification,” *European journal of echocardiography*, vol. 7, no. 2, pp. 79–108, 2006.
- [112] A. Ulloa, “Disiml: deep learning library for tabular, series, and video data.” <https://github.com/alvarouc/disiml>, 2019.