Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation[*]

Mahdi Tabassian[1], Imran Sunderji[2], Tamas Erdei[2],
Sergio Sanchez-Martinez[3], Anna Degiovanni[4], Paolo Marino[4],
Alan G. Fraser[1,2], and Jan D'hooge[1]
mahdi.tabassian@kuleuven.be

[1] Cardiovascular Imaging and Dynamics, Department of Cardiovascular Sciences,
KU Leuven, Leuven, Belgium
[2] Wales Heart Research Institute, Cardiff University, Cardiff, United Kingdom
[3] Department of Information and Communication Technologies,
Universitat Pompeu Fabra, Barcelona, Spain
[4] Department of Cardiology, University of Eastern Piedmont, Novara, Italy

Background Stress testing helps diagnose heart failure with preserved ejection fraction (HFpEF), but there are no established criteria for quantifying left ventricular (LV) functional reserve. The aim of this study was to investigate whether comprehensive analysis of the timing and amplitude of LV long-axis myocardial motion and deformation throughout the cardiac cycle during rest and stress can provide more informative criteria than standard measurements.

Methods Velocity, strain, and strain rate traces were measured from all 18 LV segments by echocardiographic myocardial velocity imaging at rest and during semisupine bicycle exercise in 100 subjects aged 69 ± 7 years, including patients with HFpEF and healthy, hypertensive, and breathless control subjects. A machine-learning algorithm, composed of an unsupervised statistical method and a supervised classifier, was used to model spatiotemporal patterns of the traces and compare the predicted labels with the clinical diagnoses.

Results The learned strain rate parameters gave the highest accuracy for allocating subjects into the four groups (overall, 57%; for patients with HFpEF, 81%), and into two classes (asymptomatic vs symptomatic; area under the curve, 0.89; accuracy, 85%; sensitivity, 86%; specificity, 82%). Machine learning of strain rate, compared with standard measurements, gave the greatest improvement in accuracy for the two-class task (+23%,

P < .0001), compared with +11% (P < .0001) using velocity and +4% (P < .05) using strain. Strain rate was also best at predicting 6-min walk distance as an independent reference criterion.

Conclusions Machine learning of spatiotemporal variations of LV strain rate during rest and exercise could be used to identify patients with HFpEF and to provide an objective basis for diagnostic classification.

Keywords

HFpEF, Diastolic stress test, Strain rate, Spatiotemporal-rest-exercise modeling, Machine learning

Highlights

- A machine-learning algorithm is proposed to diagnose HFpEF.

- Spatiotemporal patterns of echocardiographic deformation curves were analyzed.

- The algorithm's ability to identify subjects of four clinical categories was tested.

- Analyzing rest and exercise strain rate data led to the best classification accuracy.

- Machine learning of strain rate is useful for identifying HFpEF objectively.

Patients with heart failure with preserved ejection fraction (HFpEF) may have multiple pathophysiologic mechanisms for their disease, including global diastolic [1] and regional systolic dysfunction [2–4]. They can also have left ventricular (LV) mechanical dyssynchrony [5, 6] that may be revealed only during stress testing [4]. Both spatial heterogeneity and temporal dyssynchrony of function may contribute to impaired diastolic filling in patients with HFpEF, but subtle differences between regions and changes during stress can be difficult to recognize visually, because of the limited temporal resolution of human vision [7]. Two-dimensional visualization of cardiac images makes it difficult to detect three-dimensional spatial relationships, and time-separated recordings at rest and during stress hamper the detection of stress-induced effects. A complete formal analysis of myocardial velocities and deformation recorded at high frame rates might reveal much more.

Segmental velocity or deformation curves can be interpreted visually, or, for a more objective assessment, a limited number of characteristic variables (such as end-systolic strain) can be extracted from the curves and used to describe spatiotemporal interactions [8]. Examples include calculating the SDs of times to segmental peak systolic strain, to assess LV mechanical dyssynchrony during systole [6], and of times to segmental peak early diastolic velocity, to study the effects of dyssynchronous relaxation on filling [9]. Such approaches, however, do not use all the available data. Because the SDs of timings take no account of the spatial distribution of those timings, important diagnostic information might be lost.

It may now be possible to overcome these limitations by using machine learning to reveal patterns in functional traces obtained during the whole cardiac cycle [10–14]. We reported recently that such a learning scheme applied to velocity profiles acquired from the basal septum and the basal lateral wall at rest and during submaximal exercise improved the characterization of HFpEF [10, 11]. This new approach has not yet been applied to analyze function in all myocardial segments or to investigate the diagnostic utility of deformation imaging during stress.

We designed this study to test the hypothesis that patients with HFpEF have temporal and spatial variations in regional myocardial function and functional reserve that contribute to impaired global function. We used machine learning to analyze the spatiotemporal characteristics of velocity, strain, and strain rate traces acquired from all 18 segments of the left ventricle at rest and during submaximal exercise in four groups of subjects.

Methods

Study Population

One hundred subjects aged 69 ± 7 years were recruited for this prospective observational study, at two participating centers in the MEtabolic Road to DIAstolic Heart Failure (MEDIA) project [15]: the University Hospital of Wales and the Wales Heart Research Institute (Cardiff, United Kingdom) and the Scuola di Medicina of Eastern Piedmont University (Novara, Italy). At each institution, the investigators obtained ethical approval for the study from the ethics committee, and all subjects gave written informed consent.

There were 33 patients with diagnoses of HFpEF according to the consensus criteria of the European Society of Cardiology [1] and 67 control subjects subdivided into three groups. Patients with HFpEF had signs or symptoms of heart failure, preserved LV global systolic function by echocardiography (ejection fraction > 50% and LV end-diastolic volume index $< 97\ mL/m^2$), and evidence of LV diastolic dysfunction. Twenty-five breathless subjects who did not meet these criteria, and who had no respiratory disease, were identified as breathless control subjects. Asymptomatic subjects with no cardiovascular history were recruited as healthy control subjects (n = 23). Asymptomatic volunteers with no cardiovascular histories or findings apart from mildly elevated blood pressure (systolic > 140 mm Hg and/or diastolic > 90 mm Hg) were categorized as asymptomatic hypertensive control subjects (n = 19); this group included a few subjects on medical treatment. The three control groups and the subjects with HFpEF were recruited to study older subjects across a spectrum from good health to established disease. It was anticipated that the asymptomatic hypertensive control subjects would have relatively normal exercise responses, while breathless control subjects might have impaired responses that were less in magnitude than those observed in patients with HFpEF.

Exclusion criteria for all groups included any severe respiratory disease, any known coronary artery disease including myocardial infarction or revascularization, and any cerebrovascular disease or stroke within the previous 3 months. Subjects were also excluded if their echocardiographic images were not of diagnostic quality or if they were unable to perform the exercise protocol.

Echocardiography

All subjects underwent a stress test on a semisupine bicycle according to the Cardiff–MEtabolic Road to DIAstolic Heart Failure protocol, as described previously [8]. Echocardiography was performed using Vivid E9 echocardiographic machines with M4S transducers (GE Vingmed Ultrasound, Horten, Norway). Baseline images were acquired from subjects after a period of initial rest on the bicycle for 5 to 10 min, and then subjects were exercised according to a ramped protocol, starting at an initial workload of 15 W for 1 min and increasing in increments of 5 W every minute until a heart rate of 100 to 110 beats/min was achieved (or the patient developed symptoms at a lower rate). The workload was then maintained constant while exercise echocardiographic imaging was performed before fusion of the early and atrial phases of diastolic filling.

Three-beat color myocardial velocity loops of apical four-chamber, two-chamber, and long-axis images were acquired at a frame rate of 189 30 Hz and stored digitally. All images were acquired at passive end-expiration to minimize effects of translation of the heart. Offline analysis was performed using EchoPAC version 113 (GE Vingmed Ultrasound). Each myocardial wall was divided into three equal segments. A sample size (region of interest) of 1×10 mm was placed in the middle of each segment, with the first placed approximately 10 mm above the mitral annulus. Thus, velocity, strain, and strain rate curves were obtained from all 18 myocardial segments [16] at rest and during submaximal exercise. To compensate for any drift, the EchoPAC software automatically returned the deformation curves back to the baseline at the end of each beat. The derivation and clinical utility of Doppler-based deformation imaging have been reviewed elsewhere [17, 18].

Machine Learning

The overall structure of the machine-learning framework is illustrated in Figure 1. The three main phases were preprocessing of the data, statistical modeling, and automatic classification.

Preprocessing

To adjust for differences between patients in the number of temporal samples of the extracted curves, which can be due to differences in frame rate and/or heart rate, we performed a temporal alignment independently for the rest and the exercise data on the basis of linear interpolation (Figure 1A and B). After the temporal alignment, all segmental curves acquired at rest had 208 time points, while all those acquired during exercise had 123. More details are given in the Appendix.

Noisy or artifactual curves were omitted from the analysis by applying a set of quality control indexes that are described in the Appendix (see Supplemental Figure 1 for examples of artifactual strain rate curves). Some traces were missing because a segment could not be imaged or because a meaningful trace could not be extracted from that segment during postprocessing. The preprocessing phase therefore included a "data imputation" step (Figure 1C and D), during which the missing and omitted artifactual traces were estimated through a learning method called KNNimpute, as previously described [19].
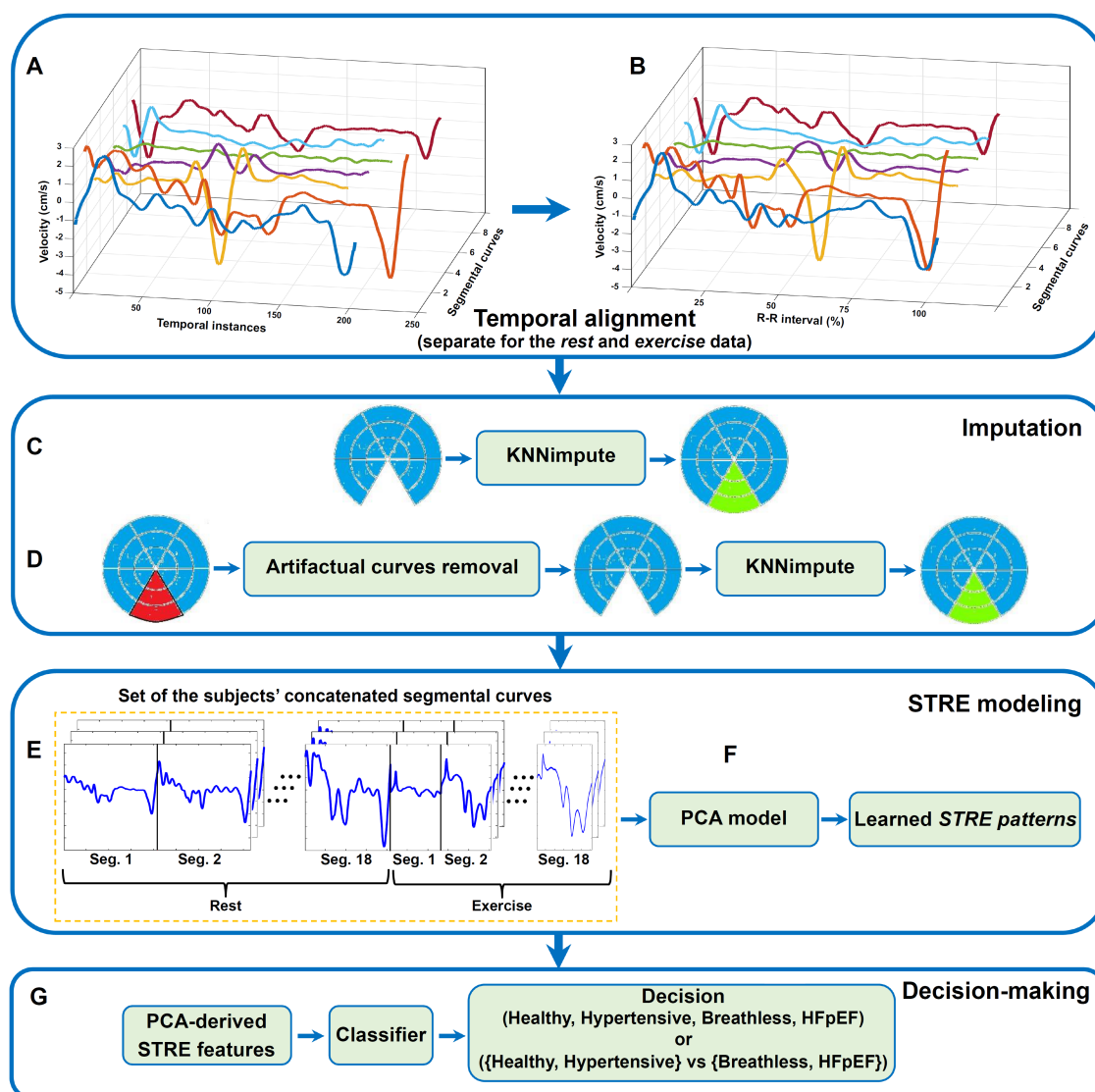
Figure 1: Overall structure of the machine-learning framework. After performing the temporal alignment separately for data obtained at rest (A) and during exercise (B), the KNNimpute approach was adopted to estimate the missing (C) or artifactual (D) curves. Each subject's segmental curves acquired at rest and during exercise were then concatenated, and the obtained set of data for the training subjects (E) was given to the PCA model to learn their STRE patterns (F). Finally, the computed STRE features of the training and testing subjects were applied respectively to build a classifier and to make decisions about the category of the subjects (G). The decision-making phase was performed twice, to classify the subjects into one of the four or two categories that were considered in this study. This procedure was carried out independently for the velocity, strain, and strain rate data.

Statistical Modeling for Spatiotemporal-Rest-Exercise Representation

After preprocessing, the temporal characteristics of segmental velocities and deformation throughout the whole cardiac cycle, as well as spatial associations at rest and during exercise, were statistically modeled. Firstly, all 36 segmental curves of each parameter that were acquired from each subject (i.e., 18 at rest and 18 during exercise) were concatenated in a fixed order on the basis of the 18-segment model of the left ventricle16 to form a "spatiotemporal-rest-exercise" (STRE) representation of LV function with 5,958 samples (Figure 1E). Principal-component analysis (PCA) [20] was then used to learn the major patterns of variation, called principal components (PCs), of the STRE data of all subjects regardless of their diagnostic categories (i.e., "unsupervised learning"; Figure 1F). We refer to the parameters learned by PCA as the "STRE feature set". We built three independent PCA models using the velocity, strain, and strain rate STRE data sets. In summary, this process allows simultaneous modeling of the behavior of all segments from the traces of the subjects acquired at rest and during exercise in a compact manner.

For comparison, we also performed modeling on the basis of PCA using either the traces obtained at rest or, independently, the traces extracted from the exercise studies. Those results are presented in the Appendix.

Automatic Classification

Three independent classifiers were constructed with the STRE feature sets of the velocity, strain, and strain rate data (Figure 1G) to categorize the subjects automatically into one of four groups—healthy control subjects, hypertensive control subjects, breathless control subjects, or patients with HFpEF—in a "supervised learning" process. In this study we used the distance-weighted -nearest-neighbor (DWKNN) method [21] as the classifier, given its ability to solve classification problems with limited training samples and to generate a probability (which necessitates using at least three KNNs [21]), not merely a categorical value, representing the likelihood that a given test subject lies within each group. In addition to this four-class task, a two-class problem was also solved, allocating the subjects by machine learning into a combined class of healthy and hypertensive subjects (i.e., all asymptomatic subjects) or into a second class including all breathless subjects and patients with HFpEF (i.e., all symptomatic subjects). Our hypothesis was that asymptomatic control subjects with mild hypertension would behave similarly to healthy control subjects, while breathless subjects would be similar to, but perhaps less severely affected than, patients fulfilling the formal criteria for HFpEF. We compared the performance of the two-class machine learning to the four-class discrimination.

More technical information about the implementation of the proposed machine-learning framework is given in the Appendix and Supplemental Figure 2.

Diagnostic Comparisons

To test the diagnostic impact of using all available mechanical data throughout the whole cardiac cycle against that of using only selected parameters as would be obtained using conventional statistical analysis, we extracted a set of markers from the segmental traces acquired at rest and during exercise in each of the 100 subjects in this study. These extracted markers were as follows:

- for velocity, (1) peak systolic amplitude, (2) time to peak systolic velocity, (3) the SD of segmental times to peak systolic velocity, and (4) peak early diastolic amplitude;

- for strain, (1) amplitude at end-systole, (2) when present, postsystolic deformation, and (3) the amplitude of diastolic strain at the end of the first third of the duration of diastole [22–24]; and

- for strain rate, (1) peak systolic amplitude, (2) peak early diastolic amplitude, (3) peak late diastolic amplitude during atrial contraction, (4) time to peak systolic strain rate, and (5) time to peak early diastolic strain rate.

The peak systolic values were extracted from the curves during the ejection phase (i.e., from aortic valve opening until aortic valve closure) by excluding electromechanical coupling and isovolumetric contraction phases [25]. To compute the peak early diastolic values, the period between mitral valve opening and the onset of the P wave on the electrocardiogram was considered, while the peak late diastolic values were calculated between the onset of the P wave and mitral valve closure.

For each subject and each type of marker, the values extracted from the segmental curves were concatenated to form a vector of markers (e.g., 36 values of peak systolic strain rate extracted from the curves acquired at rest and during exercise). This resulted in 12 different sets of markers (four sets for velocity, three sets for strain, and five sets for strain rate). The sets of markers extracted from each type of curves were also concatenated ("all markers"), forming a long vector of parameters (e.g., 180 parameters for strain rate by concatenating five sets of markers each of size 36). In total, therefore, we used 15 sets of markers, which we named "conventional feature sets," to build 15 independent DWKNN classifiers. Similar to the classification based on the STRE features, outputs of the classifiers built with the conventional features were also used to solve the two-class problem of discriminating healthy, hypertensive versus breathless, HFpEF subjects.

In addition, we compared the performance of our proposed machine-learning framework for identifying to which of the four predefined diagnostic groups each subject belonged against independent DWKNN classifiers built with some baseline echocardiographic and clinical measurements. The examined measurements were LV end-diastolic volume index, LV end-systolic volume index, left atrial volume index, and 6-min walk distance. We also concatenated these four measurements and built another classifier with the obtained vector.

In an independent set of experiments, we tested the utility of analyzing the complex spatiotemporal patterns of regional myocardial function, to predict exercise capacity. Hereto, we assessed whether the proposed learning framework is suitable for predicting diagnostic classes other than those that are already established (i.e., the four clinical categories included in the present study). We tested if independent analysis of the rest and submaximal exercise data, as well as combined rest-exercise data (i.e., STRE), on the basis of PCA could classify the subjects into binary groups of good versus poor exercise performance, using three thresholds (350, 400, and 450 m) according to their 6-min walk distance.

Statistical Analysis

As is common in machine learning [26], the subjects were randomly split into a "training set" for building the classifier and a "testing set" for evaluating its performance. The training set consisted of 64 subjects, whereby the same number of subjects was randomly taken from each of the four groups (i.e., 16 subjects per group), to ensure that all groups had the same chance in the classification phase, and the testing set contained the remaining 36 subjects. This process was repeated 100 times, and the average classification accuracy was calculated; it is reported as a mean value with its SD. Using this process, we were able to find the optimal number of PCs (i.e., STRE features) and nearest neighbors for DWKNN, and we ensured that all subjects were used in the training and testing phases and that the learning algorithm was not biased toward a subset of the subjects. A more detailed explanation of the process of training and testing the machine-learning algorithm is provided in the Appendix.

In the same manner, subjects with good and poor exercise tolerance (defined as more or less than the threshold 6-min walk distance) were randomly divided into training and testing sets such that the training set had the same number of subjects from the two groups. Similarly to the previous experiments, DWKNN was used as classifier, and the process of random sampling of the subjects was repeated 100 times.

To compare the classification results obtained with the different types of feature sets, the Mann-Whitney U test was used. Data processing and statistical analyses were performed in MATLAB version 2015b (The MathWorks, Natick, MA).

Results

The characteristics of the subjects included in this study are given in Table 1. There were 43 men and 57 women, and the mean age of all subjects was 69 ± 7 years.

The diagnostic categories assigned by the classifiers that were trained with the STRE features are compared with the clinical diagnoses in Table 2. The classifier that performed best was built on the strain rate data; it performed well in identifying subjects with HFpEF (accuracy, 81%) and otherwise classified them mainly as breathless subjects (12%). This classifier also worked moderately well when analyzing the breathless and healthy subjects, with accuracy rates of 53% and 51%, respectively. It had difficulties in differentiating breathless subjects from patients with HFpEF ("misclassification," 26%) and healthy from hypertensive control subjects ("misclassification," 29%). The velocity classifier had lower performance than that of strain rate, but its predictive pattern was similar, except that it failed to identify a majority of the hypertensive subjects, who were assigned almost equally to all four groups. The strain classifier yielded the highest performance in identifying the hypertensive subjects (although still only 49% concordance with the clinical diagnosis) but at the cost of poorer performance on the other groups. The optimal numbers of PCs used for building the STRE feature sets for the velocity, strain, and strain rate curves were 50, 10, and 25, respectively. Using these numbers of PCs, > 90% of the data variation was captured.

Table 1: Characteristics of control subjects and patients.

	Healthy subjects (n=23)	Hypertensive subjects (n=19)	p-value for healthy vs. hypertensive	Breathless subjects (n=25)	HFpEF Patients (n=33)	p-value for breathless vs. HFpEF	All subjects
Age (y)	66.1 ± 3.6	67 ± 9	0.71	68 ± 4.7	73 ± 7.6	0.02	69.1 ± 7.1
Female, N (%)	12 (52)	6 (32)	0.19	18 (72)	21 (91)	0.51	57(57)
Body mass index (kg/m^2)	27.2 ± 4.1	26.1 ± 3.4	0.63	28 ± 4.9	31.73 ± 5.7	0.01	28.7 ± 5.2
6-minute walk test (m)	515.2 ± 75	532.7 ± 94.6	0.4	446.5 ± 124.3	349.3 ± 102.4	0.001	449.6 ± 124.7
Biomarkers							
NT-proBNP (ng/L)*	73.8 ± 58	85.5 ± 62.4	0.44	88.2 ± 59.4	175.2 ± 151.3	0.26	118.2 ± 114.2
Creatinine (μmol/L)*	74.4 ± 9	71 ± 8	0.34	79.5 ± 19	82.2 ± 24	0.88	77.9 ± 18.4
Hemodynamic data							
Heart rate (bpm), rest	65 ± 7	61 ± 17	0.77	70 ± 11	65 ± 10	0.1	66 ± 12
Heart rate (bpm), peak exercise	132 ± 9	123 ± 16	0.03	118 ± 16	101 ± 12	< 0.0001	117 ± 18
Systolic blood pressure (mmHg), rest	119 ± 18	134 ± 15	0.008	134 ± 16	135 ± 17	0.81	131 ± 18
Diastolic blood pressure (mmHg), rest	70 ± 11	76 ± 13	0.22	75 ± 10	73 ± 17	0.97	73 ± 13
Systolic blood pressure (mmHg), peak exercise	181 ± 18	192 ± 26	0.14	177 ± 20	179 ± 27	0.59	182 ± 23
Diastolic blood pressure (mmHg), peak exercise	78 ± 16	86 ± 13	0.16	86 ± 13	79 ± 18	0.08	82 ± 16
Echocardiography							
LV end diastolic volume index (ml/m^2)	41.1 ± 7.5	42.1 ± 9.4	0.84	38.9 ± 8.5	42.9 ± 12.7	0.26	41.3 ± 10
LV end systolic volume index (ml/m^2)	15 ± 2.5	14.8 ± 3.3	0.7	15 ± 5	16.4 ± 5.7	0.27	15.4 ± 4.5
LV ejection fraction (%)	63.2 ± 3.8	64.7 ± 2.6	0.12	63 ± 5.7	62.1 ± 4.8	0.78	62.8 ± 4.6
LV mass index (g/m^2)	74.5 ± 16.9	98.6 ± 28.4	0.003	97.8 ± 33.4	109 ± 24.4	0.06	96.3 ± 29
E/A ratio, rest	0.86 ± 0.18	0.83 ± 0.3	0.16	0.81 ± 0.29	0.79 ± 0.17	0.45	0.82 ± 0.2
Deceleration time (ms), rest	244.3 ± 37.6	240.4 ± 59	0.3	235.6 ± 60	253.6 ± 56.3	0.42	244.5 ± 53.7
Ard-Ad (ms)**	-7 ± 17.5	-6.6 ± 30	0.95	-10.3 ± 17.9	3.4 ± 25	0.135	-4.3 ± 22.9
E/e' ratio, rest	7.23 ± 2.2	8.6 ± 2.8	0.05	9 ± 3.7	10.8 ± 3.6	0.03	9.1 ± 3.4
E/e' ratio, exercise	7.8 ± 2.5	8.2 ± 1.9	0.39	8.3 ± 2.2	10.4 ± 2.9	0.004	8.8 ± 3.4
LA volume index (ml/m^2)	27.1 ± 8.1	33.14 ± 8.8	0.03	32.5 ± 11.8	33.5 ± 10	0.24	31.7 ± 10.1

* Based on measurements from Cardiff subjects and HFpEF patients from Novara

** Based on measurements from Cardiff subjects

Table 2: Comparison of diagnoses predicted by the machine learning algorithm for spatio-temporal changes, and the diagnostic categories assigned using consensus recommendations

Diagnosis predicted by machine learning	Clinical labels pre-assigned by recommended diagnostic criteria			
	Healthy	Hypertensive	Breathless	HFpEF
Using velocity				
Healthy	**48**	31	19	6
Hypertensive	16	**25**	8	7
Breathless	12	24	**46**	14
HFpEF	24	20	27	**73**
Using strain				
Healthy	**29**	22	34	17
Hypertensive	29	**49**	10	20
Breathless	21	14	**30**	33
HFpEF	21	15	26	**30**
Using strain rate				
Healthy	**51**	19	12	4
Hypertensive	29	**41**	9	3
Breathless	8	23	**53**	12
HFpEF	12	17	26	**81**

Each reported value is the mean result of 100 classification launches. The results are given as percentages. The diagonal entries (in bold) show the concordant classification rates.
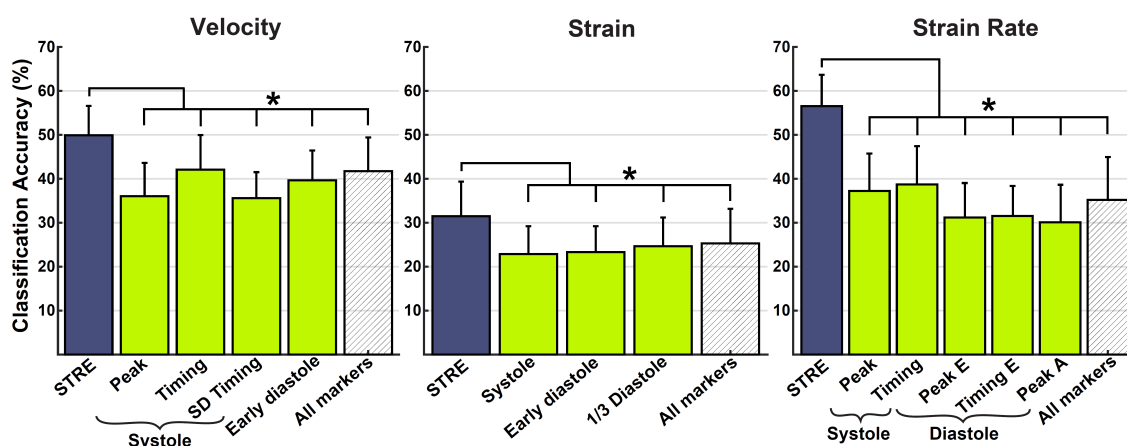
Figure 2: Diagnostic accuracy of STRE patterns compared with conventional features of segmental function. Classification accuracy rates (mean ± SD) are shown for the STRE and conventional feature sets, separately for velocity, strain, and strain rate data, for the four-class case. ∗ P < .0001. Velocity measurements were peak systolic velocity, time to peak systolic velocity (Timing), the SD of times to peak systolic velocities (SD Timing), and early diastolic velocity. For strain, peak amplitude during systole, and during early diastole and the amplitude after the first third of diastole (1/3 diastole) were computed. Strain rate measurements were peak systolic amplitude, time to the systolic peak (Timing), peak early diastolic amplitude (Peak E) and timing (Timing E), and peak amplitude during atrial filling (Peak A). "All markers" refers to the results obtained using a concatenated sequence of all the standard measurements.

Automatic Classification Compared with Clinical Diagnoses

The diagnostic accuracy of machine learning, using the features extracted by analysis of the whole signal throughout the cardiac cycle at rest and during exercise in all 18 segments ("STRE"), is compared with the performance of the conventional features in Figure 2. For each variable (velocity, strain, and strain rate), STRE was more accurate for classifying the subjects into the four groups determined using conventional diagnostic labels (P < .0001). The overall accuracy of STRE was highest using strain rate (57%), compared with 50% using velocity and 31% using strain data. Among the extracted conventional features, the highest diagnostic accuracy was obtained using the time to peak systolic velocity (42%).

The classification performance of the machine-learning framework trained with the STRE features of the strain rate curves is contrasted with the results obtained with the baseline echocardiographic and clinical measurements in Figure 3. The STRE features yielded significantly better results than the individual and combined measurements (P < .0001) for classifying the subjects into four groups. The results obtained with left atrial volume index were better than the other measurements (51%), but the classifier built with this parameter was biased toward the breathless control subjects and patients with HFpEF and failed to correctly classify the healthy and hypertensive subjects.
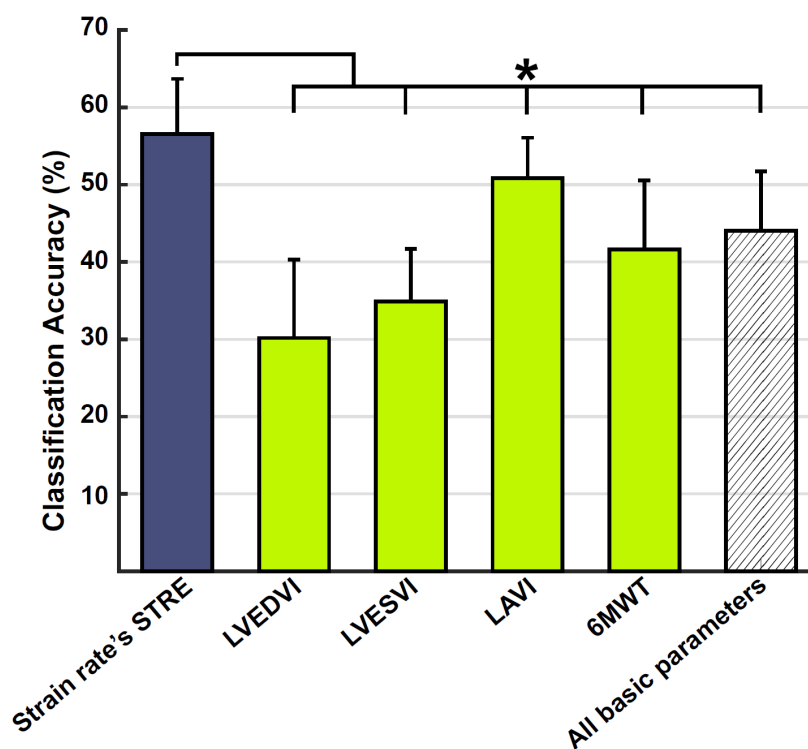
Figure 3: Diagnostic accuracy of STRE patterns compared with left heart volumes and with exercise capacity. Accuracy rates (mean ± SD) for allocating subjects to four classes, obtained using the STRE features of the strain rate curves, compared with baseline echocardiographic measurements of indexed LV end-diastolic volume (LVEDVI) and indexed LV end-systolic volume (LVESVI), indexed left atrial volume (LAVI), and the 6-min walk test (6MWT) distance. Results are also shown for a feature vector consisting of all the baseline measurements ("All basic parameters") and the 6MWT results. ∗ P < .0001.

Discriminating Symptomatic from Asymptomatic Subjects

Machine learning using STRE data was compared with extracted conventional feature sets for their ability to separate subjects into two classes, namely, asymptomatic versus symptomatic subjects. The results are illustrated as receiver operating characteristic (ROC) analyses in Figure 4. For each echocardiographic variable that was studied, the STRE feature set had a larger area under the curve (AUC) than the conventional feature sets. The highest AUCs were obtained with strain rate STRE (0.89) and velocity STRE features (0.84). The conventional features extracted from the velocity curves resulted in larger AUCs than those obtained with the other conventional features.

The accuracy, sensitivity, and specificity of each feature set were calculated using optimal cutoff values obtained from the results of the ROC analyses. The results are presented in Figure 5 and compared with the performance of the four-class case. For each classifier, the optimal cutoff value was selected by plotting its sensitivity and specificity curves obtained with examining different cutoffs and taking a value that led to similar sensitivity and specificity results. The highest average accuracy, sensitivity, and specificity were obtained using strain rate STRE, at 85%, 86%, and 82%, respectively. The
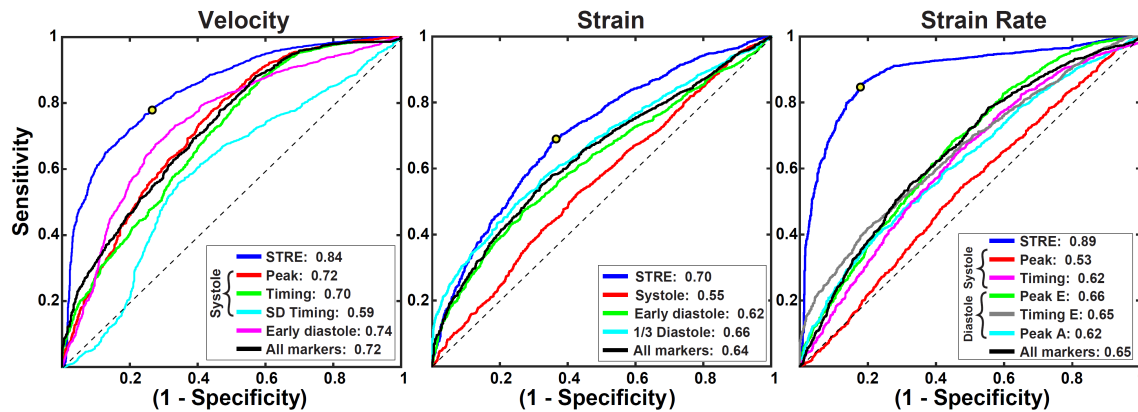
Figure 4: ROC curves comparing the diagnostic performance of STRE against conventional features. ROC analyses for the two-class classification achieved with STRE and with the conventional features of the velocity, strain, and strain rate data. The optimal cutoff values for the STRE features are demonstrated with a circle.

peak early diastolic velocity marker had higher diagnostic performance than the other conventional features (average accuracy, sensitivity, and specificity of 70%, 70%, and 69%, respectively). All classifications into two groups were more accurate than classifications into four groups (P < .0001).

In addition to the ROC analyses, H measures [27] of the classifiers, as an alternative to AUC, were computed using an R package called hmeasure (https://cran.r-project.org/web/packages/hmeasure/index.html), and the results are presented in Supplemental Table 1.

Prediction of Exercise Capacity

Six-minute walk distances in this study ranged from a minimum of 180 m in one of the subjects with HFpEF to a maximum of 690 m in one of the healthy control subjects, but there was considerable overlap between groups. We used ROC analysis to assess the ability of machine learning to discriminate subjects according to a threshold value for their 6-min walk distance of 400 m (Figure 6). For velocity, strain, and strain rate, analyses using all the features at rest and on exercise (STRE) gave slightly larger AUCs. The highest values were obtained using features derived by machine learning from the strain rate curves acquired during exercise (0.83) and during rest and exercise (STRE; 0.86).

The AUCs, H values, and average accuracy, sensitivity, and specificity of the different feature sets for three examined thresholds are listed in Supplemental Table 2, Supplemental Table 3, Supplemental Table 4.

Discussion

This study suggests that (1) analysis of the complex spatiotemporal patterns of regional myocardial mechanical function at rest and during exercise is a promising method for
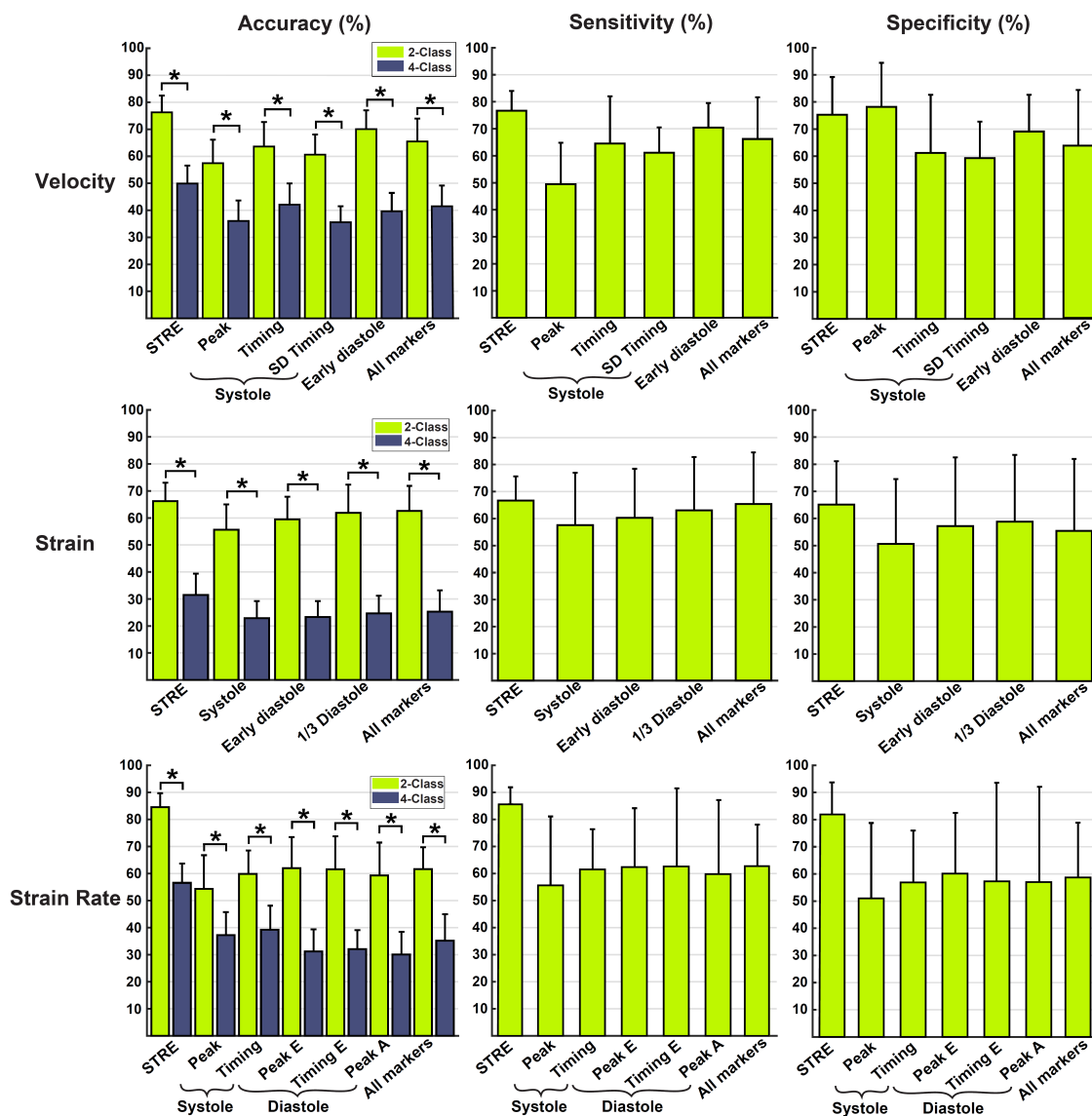
Figure 5: Comparison of diagnostic performance by clinical labels or by symptomatic status. Accuracy, sensitivity, and specificity (%; mean $\pm$ SD) of the optimal cutoff values for velocity, strain, and strain rate derived from the ROC analyses presented in Figure 4, for identifying two classes (breathless or asymptomatic), compared with the accuracy rates for identifying four classes (healthy, asymptomatic hypertensive, breathless, and breathless with HFpEF). Abbreviations as in Figure 2. $*$ P < .0001.
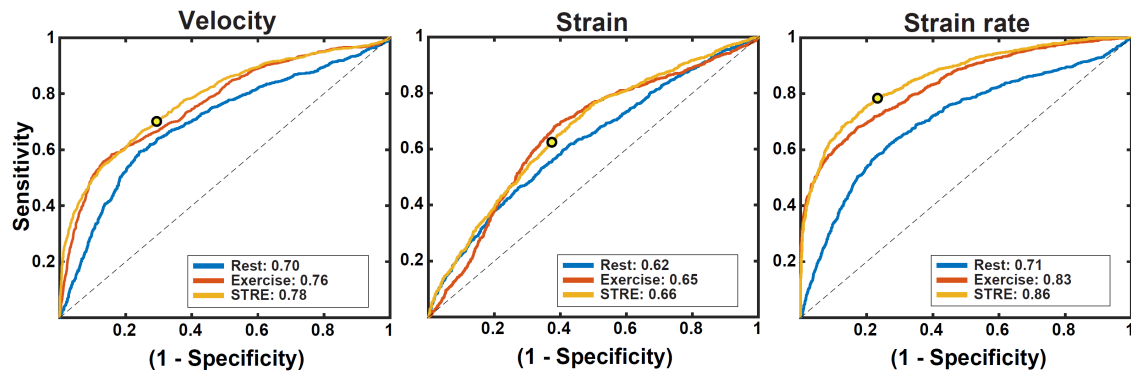
Figure 6: ROC curves for machine learning to predict exercise capacity. ROC analyses summarizing the classification outcomes obtained using the spatiotemporal data at rest, during exercise, and combining rest and exercise data (STRE), to predict the 6-min walk distance applying a threshold of 400 m. The optimal cut points for the STRE features are demonstrated with a circle.

identifying HFpEF, and (2) the strain rate of longitudinal deformation has higher diagnostic power than either myocardial velocity or segmental strain.

Impaired exercise capacity in HFpEF may be caused by different mechanisms in subjects in the early or later stages of its natural history. We tested the hypothesis that patients with HFpEF have more temporal and spatial variations in LV regional myocardial diastolic and/or systolic function than healthy subjects, because such variations would reduce early diastolic filling by inefficient relaxation and reduce cardiac output by causing dyssynchronous contraction. To extract as much information as possible from all LV segments both at rest and during exercise, we used machine learning to analyze the spatiotemporal patterns of regional myocardial mechanics.

The results presented in the present study are in agreement with our recent investigation [12] demonstrating that modeling spatiotemporal characteristics of baseline strain rate curves using machine learning is more efficient than using the same model built with strain curves for identifying myocardial infarction. In the present study, we extended our previous model by analyzing the baseline and exercise data simultaneously and showed that such an analysis is more useful than modeling the baseline or exercise data independently (see Supplemental Figure 3).

Diagnostic Classification

HFpEF is a clinical syndrome that can result from many etiologies, and any diagnostic study is limited by the lack of definitive reference criteria. To recruit patients with HFpEF, we applied the consensus recommendations [1], but it is recognized that these have some limitations [28–30]. Because they do not include any diagnostic information obtained during exercise, some subjects without significant respiratory disease, coronary artery disease, or heart valve disease, who have breathlessness from a cardiac cause but relatively normal function at rest, may be diagnosed as normal. Other subjects may be diagnosed as having HFpEF because of atrial fibrillation and some echocardiographic findings that are common with aging. In addition to an age-matched control group of

healthy asymptomatic subjects, therefore, we also recruited two intermediate control groups. An asymptomatic mildly hypertensive group was included because there is little consensus on how to define healthy older people as control subjects, in particular whether they should have any risk factors apart from age. The breathless control group was included because we suspected that their functional capacity would be little different from the patients who fulfilled the formal diagnostic criteria for HFpEF. We envisaged that taken together, these four groups would include subjects across a wide spectrum from health to disease.

We used machine learning first to classify subjects into the four groups as defined by the clinical diagnostic criteria. Performing this task purely at random would result in a 25% overall accuracy, assuming that each subject has a likelihood of one in four to be in a particular category. The STRE feature set performed substantially better, particularly when identifying patients with HFpEF, since it classified 81% accurately based only on the characteristics of their regional myocardial function and deformation.

Unsurprisingly, the highest probability of error in classifying the breathless subjects was due to confusion with the patients with HFpEF, whereby 26% of the breathless subjects—used in the testing set during the 100 runs of the algorithm—were misclassified as having HFpEF. One can question whether these subjects were truly misclassified on the basis of their LV mechanical characteristics or if they had been attributed an incorrect clinical label. This potential misclassification effect was mitigated when machine learning was used to divide the subjects into symptomatic versus asymptomatic subjects (i.e., a two-class classification problem); this gave higher overall accuracies, up to 85% (see Figure 5).

An important question is whether the observed classification performance was achieved because the full spatiotemporal myocardial mechanical information was considered or whether the essential information could be captured by standard markers derived from these data sets. We found that the STRE patterns showed a clear advantage over the conventional feature sets and the baseline echocardiographic and clinical measurements, thereby demonstrating that current measurements such as the peak amplitudes of systolic and early diastolic velocity or strain rate, or indexed LV end-diastolic and end-systolic volumes, do not exploit all available diagnostic information. The number of STRE parameters used (e.g., 25 for strain rate) was smaller than the number of parameters of the concatenated segmental curves (i.e., 5,958) and also smaller than the number of concatenated conventional echocardiographic markers (e.g., 180 for strain rate), implying that the statistical model used (i.e., PCA) was computationally efficient. Interestingly, classification performance was better ($P < .0001$) when using velocity or strain rate data acquired during exercise (Supplemental Figure 3, available at www.onlinejase.com) than at rest, suggesting that any LV mechanical differences among the four groups become amplified during stress. This would therefore support the use of stress testing to reveal HFpEF. In addition, when using strain rate, simultaneous modeling of both the rest and exercise data (i.e., STRE) led to better results ($P < .0001$) than modeling exercise data alone. This supports the utility of analyzing both the rest and exercise data for diagnostic classification.

In this study, we also examined the utility of machine learning of the full spatiotemporal myocardial mechanical information for classifying the subjects into good or poor exercise performance groups as defined by their 6-min walk test results. As expected,

modeling the submaximal exercise data led to better results than those obtained using the resting data because the signs and symptoms of abnormal LV function can be revealed by exercise testing. However, analyzing both the resting and exercise data (i.e., STRE) appeared to be more efficient, suggesting that resting data can also contribute to the prediction of exercise capacity.

Clinical Implications

It could reasonably be asked whether machine learning in this study added any particular value. Indeed, its outcome corresponded only moderately to diagnostic labels based on current clinical consensus statements. However, this criticism could be reversed, as one could claim that using sophisticated spatiotemporal analysis of LV mechanics did not enable separation of some patients with HFpEF from the control groups, thereby suggesting that current clinical diagnostic criteria might be suboptimal. Indeed, the analysis performed significantly better when dividing subjects into those who were asymptomatic and those who were breathless on exertion. When compared against an independent reference for exercise capacity (i.e., 6-min walk distance), our machine-learning algorithm was almost 80% accurate, even when nonphysiologic factors influencing walk distance or other determinants of functional capacity in HFpEF, such as peripheral muscle metabolism and oxygenation, were not taken into account. Cardiopulmonary stress tests with measurement of Vo2max or the VE/Vco2 slope would have been more useful but were not performed routinely in this study.

In future, the process of extracting segmental traces of LV long-axis function could be automated and used in a machine-learning algorithm that gives the probability of disease as its output. This would provide a purely objective diagnosis but with the option to reconstruct curves from the PCs that allow some pathophysiologic interpretation. The main purpose of our study was to investigate how this could be performed.

In future, the process of extracting segmental traces of LV long-axis function could be automated and used in a machine-learning algorithm that gives the probability of disease as its output. This would provide a purely objective diagnosis but with the option to reconstruct curves from the PCs that allow some pathophysiologic interpretation. The main purpose of our study was to investigate how this could be performed.

Choice of Parameter

Since the introduction of speckle-tracking as an alternative to color myocardial velocity imaging (tissue Doppler) for measuring myocardial deformation, it has been applied as a tool for measuring changes in function during stress [31]. The frame rate of speckle-tracking is much lower than myocardial velocity imaging, however, and we have demonstrated that high–frame rate acquisitions are needed to optimize the performance of machine learning. The variability of measurements of segmental strain using speckle-tracking is not trivial [32], which will also compromise comparisons of function in different segments. The main concern with using strain, however, is that it is decreased at peak stress if there is a reduction in LV volumes [33] and that is strongly load dependent [34]. In our study, strain was not useful for quantifying peak myocardial responses, which suggests that global longitudinal strain will not be a good criterion for measuring maximal stress

responses in patients with suspected HFpEF.

The best classification results were achieved using STRE features derived from segmental recordings of myocardial strain rate. These findings reinforce earlier reports that strain rate, which is a less load sensitive index, correlates well with local and global LV contractile function [34, 35] and that it is more sensitive than strain, for example to detect inducible ischemia during dobutamine stress testing [36], to detect infarcted myocardium [12], and to predict outcome in cardiac amyloidosis with normal ejection fraction [37]. Strain rate curves derived from myocardial velocity imaging can appear quite noisy but were shown to contain more diagnostic information given an appropriate analysis than the less noisy strain curves. The incorporation of PCA as the statistical model in our proposed framework ensured that the diagnostic information within the strain rate curves was preserved and the noise was removed. In fact, PCA can serve as a noise-filtering technique [20] whereby the important PCs (i.e., the first ones that were used in our experiments) modeled major patterns of the strain rate curves, and the unimportant PCs (i.e., the last ones that were ignored in our analyses) captured their noise.

One of the most sensitive indicators of myocardial diastolic dysfunction is early diastolic relaxation [38], but current commercial tools for echocardiographic analysis do not usually provide any means for measuring it on strain curves. Using machine learning, therefore, we applied the concept of the "first-third filling fraction" [22–24] but we found that in our study it did not perform better when used as a single index in identifying subjects with HFpEF.

Pathophysiologic Interpretation

Concatenated velocity and strain rate traces from a healthy control and a patient with HFpEF are displayed in Figure 7. These subjects were consistently classified correctly during the iterations of our machine-learning algorithm (see Supplemental Figure 4). The healthy subject has larger increments in strain rate during stress, particularly during early diastole and particularly in the free walls of the left ventricle rather than the septum.

It is also possible to construct traces that show variations in the PCs identified by the machine-learning algorithm as being most discriminant between subjects. Traces demonstrating variability of the first two (and therefore the most important) PCs for each of velocity, strain, and strain rate, are shown in Supplemental Figure 5, Supplemental Figure 6, Supplemental Figure 7.

The first PC of the velocity traces appears to correspond to the amplitude of peak velocities at rest and during submaximal exercise. The second is more difficult to interpret from a physiologic perspective, but there is variation in the rate of increase of early diastolic velocity at rest and in the timing of the peak systolic velocity during exercise.

The first two PCs of strain both show variations in the peak amplitude of shortening at rest, but with segmental patterns including less variation in the basal septal segments than the basal lateral segments; the former is more affected by aging and therefore may be less discriminant between healthy and diseased subjects. The first PC shows less variation in the traces during exercise, perhaps because of the load dependency of strain and its biphasic response to all stress tests.

In comparison, the first PC of strain rate shows more variation in amplitudes during exercise, both in systole and in early diastole. There are also differences between
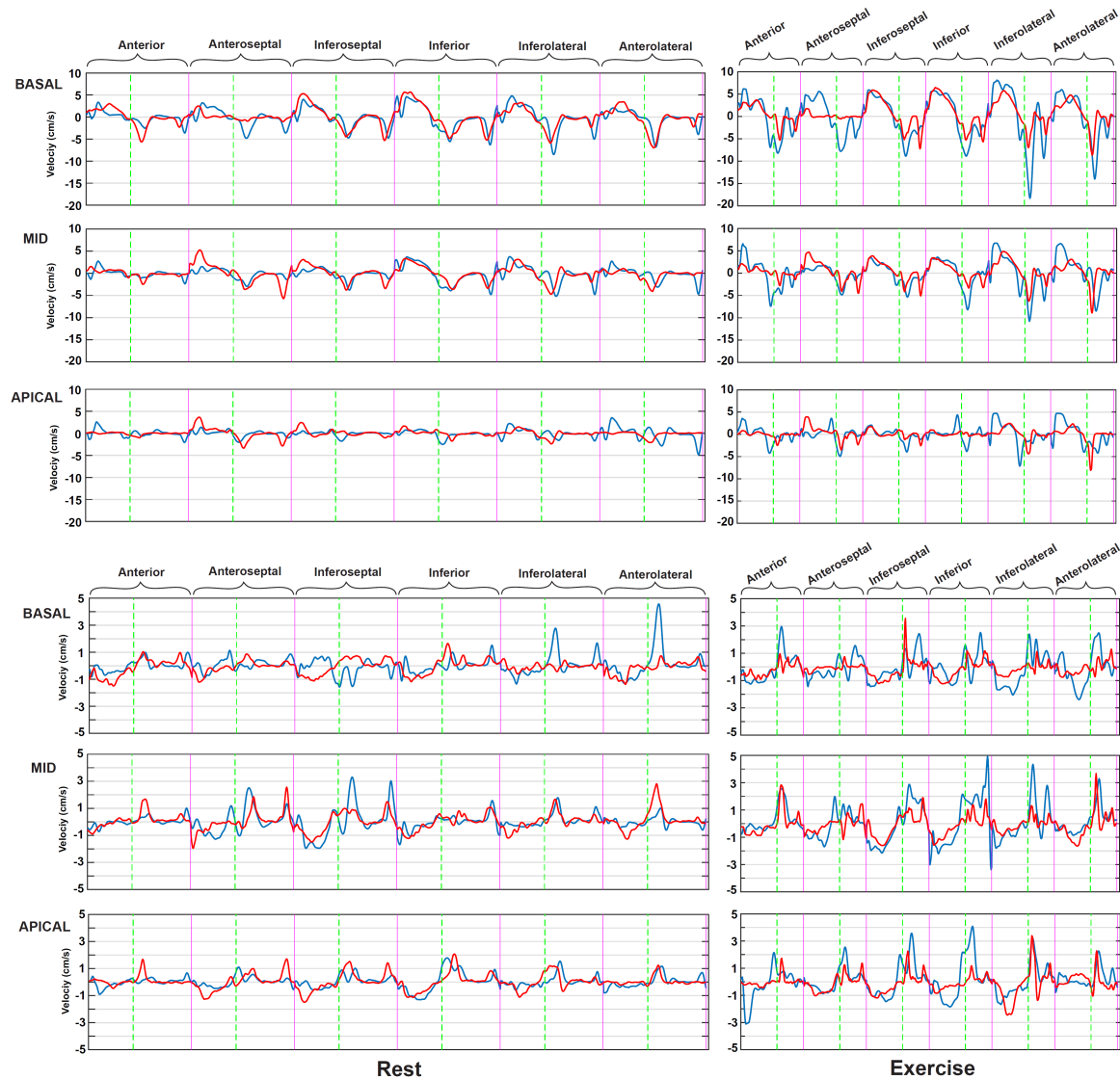
Figure 7: Myocardial function in all segments at rest and during exercise. Concatenated segmental velocity curves (top) and strain rate curves (bottom) from all 18 LV segments, in a healthy control subject (blue) and a patient with HFpEF (red) who were correctly classified by the machine learning method during all 100 repeated analyses. For each segmental curve, end-systole is indicated by a green vertical dashed line. To make it possible to appreciate variations in amplitude from rest to exercise, the scales of the vertical axes are constant, which makes some segments with very small amplitudes at rest appear almost flat.

segments, particularly at rest. The second PC of strain rate is difficult to assess, illustrating how machine learning can discern patterns for diagnosis that are not always physiologically interpretable.

Study Limitations

During the off-line analysis of the acquired images, no angle corrections were used. This might have influenced the accuracy of the measurements, but care was taken to align both LV walls in each image to the axial planes of the ultrasound beams as much as possible.

The scalar or "standard" measurements (or conventional feature sets) that we compared with the complete spatiotemporal patterns, as inputs for the machine-learning algorithm, were extracted from the digital traces using a MATLAB program. The timings that we used to define systole and diastole were global rather than regional, taken from aortic and mitral flow signals obtained at each stage of the exercise protocol. Of course, these are not synchronous with each subsequent individual beat or segmental trace. It is possible, therefore, that some of the suboptimal diagnostic performance of the conventional features could be due to inaccuracy in the timing of the phases of the cardiac cycle. There is currently no obvious solution to this problem, as simultaneous and beat-by-beat acquisition of blood flow and myocardial velocity or deformation signals is not possible. That is another reason why analysis of function throughout the cardiac cycle by machine learning may be advantageous and why it is likely to be more reproducible.

We performed an exploratory study in a small population, which may limit the accuracy of our proposed machine-learning algorithm. To account for this and to effectively test the generalization power of the learning algorithm, we ran it 100 times with randomly selected sets of training and testing subjects and then computed and reported the average performance. Our results confirm in principle that machine learning may be very informative to interpret myocardial responses to stress, but of course larger numbers of subjects should be studied prospectively to assess the efficacy of the proposed automatic classification system and to further develop it.

Conclusion

In our study, we used statistical modeling and machine learning to analyze strain rate curves from all segments of the left ventricle, acquired both at rest and during exercise. Variations in strain rate can demonstrate regional and global dysfunction as well as dyssynchrony. A machine-learning approach accomplished this complex analysis in an objective and accurate manner, so it could aid clinicians in identifying patients with HFpEF. Machine learning of the STRE patterns of segmental strain rate profiles is a promising approach that now merits further evaluation.

References

[1]   W. J. Paulus, C. Tschöpe, J. E. Sanderson, et al. "How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology". In: *Eur. Heart J.* 28.20 (2007), pp. 2539–2550.

[2]   C.-M. Yu, H. Lin, H. Yang, S.-L. Kong, Q. Zhang, and S. W.-L. Lee. "Progression of systolic abnormalities in patients with isolated diastolic heart failure and diastolic dysfunction." In: *Circulation* 105.10 (2002), pp. 1195–201.

[3]   G. Cioffi, M. Senni, L. Tarantini, et al. "Analysis of circumferential and longitudinal left ventricular systolic function in patients with non-ischemic chronic heart failure and preserved ejection fraction (from the CARRY-IN-HFpEF study)." In: *Am. J. Cardiol.* 109.3 (2012), pp. 383–9.

[4]   T. Erdei, S. Aakhus, P. Marino, W. J. Paulus, O. A. Smiseth, and A. G. Fraser. "Pathophysiological rationale and diagnostic targets for diastolic stress testing." In: *Heart* 101.17 (2015), pp. 1355–60.

[5]   J. Wang, K. M. Kurrelmeyer, G. Torre-Amione, and S. F. Nagueh. "Systolic and diastolic dyssynchrony in patients with diastolic heart failure and the effect of medical therapy". In: *J. Am. Coll. Cardiol.* 49.1 (2007), pp. 88–96.

[6]   A. B. S. Santos, E. Kraigher-Krainer, N. Bello, et al. "Left ventricular dyssynchrony in patients with heart failure and preserved ejection fraction". In: *Eur. Heart J.* 35.1 (2014), pp. 42–47.

[7]   J.-P. E. Kvitting, L. Wigström, J. M. Strotmann, and G. R. Sutherland. "How accurate is visual assessment of synchronicity in myocardial motion? An in vitro study with computer-simulated regional delay in myocardial motion: clinical implications for rest and stress echocardiography studies". In: *J Am Soc Echocardiogr* 12.9 (1999), pp. 698–705.

[8]   T. Erdei, O. A. Smiseth, P. Marino, and A. G. Fraser. "A systematic review of diastolic stress tests in heart failure with preserved ejection fraction, with proposals from the EU-FP7 MEDIA study group". In: *Eur. J. Heart Fail.* 16.12 (2014), pp. 1345–1361.

[9]   A. P.-W. Lee, J.-K. Song, G. W.-K. Yip, et al. "Importance of dynamic dyssyn-chrony in the occurrence of hypertensive heart failure with normal ejection fraction". In: *Eur. Heart J.* 31.21 (2010), pp. 2642–2649.

[10] S. Sanchez-Martinez, N. Duchateau, T. Erdei, A. G. Fraser, B. H. Bijnens, and G. Piella. "Characterization of myocardial motion patterns by unsupervised multiple kernel learning." In: *Med. Image Anal.* 35 (2017), pp. 70–82.

[11] S. Sanchez-Martinez, N. Duchateau, T. Erdei, et al. "Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction." In: *Circ. Cardiovasc. Imaging* 11.4 (2018), e007138.

[12] M. Tabassian, M. Alessandrini, L. Herbots, et al. "Machine learning of the spatio-temporal characteristics of echocardiographic deformation curves for infarct classification". In: *Int. J. Cardiovasc. Imaging* 33.8 (2017), pp. 1159–1167.

[13] M. Tabassian, M. Alessandrini, L. De Marchi, et al. "Principal Component Analysis for the Classification of Cardiac Motion Abnormalities Based on Echocardiographic Strain and Strain Rate Imaging". In: vol. 9126. The 8th International Conference on Functional Imaging and Modeling of the Heart, Lecture Notes in Computer Science. Springer, 2015, pp. 83–90.

[14] P. P. Sengupta, Y. M. Huang, M. Bansal, et al. "Cognitive machine-learning algorithm for cardiac imaging; a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy". In: *Circ. Cardiovasc. Imaging* 9.6 (2016).

[15] *MEDIA: https://cordis.europa.eu/project/rcn/96870_en.html.*

[16] R. M. Lang, L. P. Badano, V. Mor-Avi, 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." In: *J. Am. Soc. Echocardiogr.* 28.1 (2015), 1–39.e14.

[17] G. R. Sutherland, G. Di Salvo, P. Claus, J. D'hooge, and B. Bijnens. "Strain and strain rate imaging: a new clinical approach to quantifying regional myocardial function". In: *J. Am. Soc. Echocardiogr.* 17.7 (2004), pp. 788–802.

[18] J. D'hooge, A. Heimdal, F. Jamal, et al. "Regional strain and strain rate measurements by cardiac ultrasound: principles, implementation and limitations". In: *European Heart Journal-Cardiovascular Imaging* 1.3 (2000), pp. 154–170.

[19] M. Tabassian, M. Alessandrini, R. Jasaityte, L. De Marchi, G. Masetti, and J. D'hooge. "Handling missing strain (rate) curves using K-nearest neighbor imputation". In: *2016 IEEE Int. Ultrason. Symp.* IEEE, 2016, pp. 1–4.

[20] I. Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[21] S. A. Dudani. "The Distance-Weighted k-Nearest-Neighbor Rule". In: *IEEE Trans. Syst. Man. Cybern.* 6.4 (1976), pp. 325–327.

[22] I. Inouye, B. Massie, D. Loge, et al. "Abnormal left ventricular filling: an early finding in mild to moderate systemic hypertension". In: *Am J Cardiol* 53 (1964), pp. 120–126.

[23] A. Yamamoto, N. Takahashi, K. Abe, Y. Kobayashi, J. Tamai, and K. Munakata. "Regional left-ventricular diastolic wall motion assessed by a new program for ECG-gated myocardial perfusion SPECT in early-stage heart failure". In: *J. Nucl. Cardiol.* 15.3 (2008), pp. 375–382.

[24] L. A. Reduto, W. J. Wickemeyer, J. B. Young, et al. "Left ventricular diastolic performance at rest and during exercise in patients with coronary artery disease. Assessment with first-pass radionuclide angiography." In: *Circulation* 63.6 (1981), pp. 1228–37.

[25] J. D'hooge, B. Bijnens, J. Thoen, F. Van de Werf, G. Sutherland, and P. Suetens. "Echocardiographic strain and strain-rate imaging: a new tool to study regional myocardial function". In: *IEEE Trans. Med. Imaging* 21.9 (2002), pp. 1022–1030.

[26] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer New York, 2009.

[27] D. J. Hand. "Measuring classifier performance: a coherent alternative to the area under the ROC curve". In: *Mach. Learn.* 77.1 (2009), pp. 103–123.

[28] M. Penicka, J. Bartunek, H. Trakalova, et al. "Heart failure with preserved ejection fraction in outpatients with unexplained dyspnea: a pressure-volume loop analysis". In: *J. Am. Coll. Cardiol.* 55.16 (2010), pp. 1701–1710.

[29] O. F. Sharifov, C. G. Schiros, I. Aban, T. S. Denney, and H. Gupta. "Diagnostic accuracy of tissue Doppler index E/e' for evaluating left ventricular filling pressure and diastolic dysfunction/heart failure with preserved ejection fraction: a systematic review and meta-analysis." In: *J. Am. Heart Assoc.* 5.1 (2016), e002530.

[30] O. F. Sharifov and H. Gupta. "What is the evidence that the tissue Doppler index E/e' reflects left ventricular filling pressure changes after exercise or pharmacological intervention for evaluating diastolic function? a systematic review." In: *J. Am. Heart Assoc.* 6.3 (2017), e004766.

[31] V. Uusitalo, M. Luotolahti, M. Pietilä, et al. "Two-dimensional speckle-tracking during dobutamine stress echocardiography in the detection of myocardial ischemia in patients with suspected coronary artery disease." In: *J. Am. Soc. Echocardiogr.* 29.5 (2016), 470–479.e3.

[32] O. Mirea, E. D. Pagourelias, J. Duchenne, et al. "Intervendor Differences in the Accuracy of Detecting Regional Functional Abnormalities: A Report From the EACVI-ASE Strain Standardization Task Force". In: *JACC Cardiovasc. Imaging* 11.1 (2018), pp. 25–34.

[33] G. Davidavicius, M. Kowalski, R. I. Williams, et al. "Can regional strain and strain rate measurement be performed during both dobutamine and exercise echocardiography, and do regional deformation responses differ with different forms of stress testing?" In: *J. Am. Soc. Echocardiogr.* 16.4 (2003), pp. 299–308.

[34] V. Ferferieva, A. Van den Bergh, P. Claus, et al. "The relative value of strain and strain rate for defining intrinsic myocardial function". In: *Am J Physiol Heart Circ Physiol* 302.1 (2012), pp. 188–195.

[35] F. Weidemann, F. Jamal, G. R. Sutherland, et al. "Myocardial function defined by strain rate and strain during alterations in inotropic states and heart rate". In: *Am. J. Physiol. Circ. Physiol.* 283.2 (2002), H792–H799.

[36] C. Bjork Ingul, A. Stoylen, S. A. Slordahl, R. Wiseth, M. Burgess, and T. H. Marwick. "Automated Analysis of Myocardial Deformation at Dobutamine Stress Echocardiography. An Angiographic Validation". In: *J. Am. Coll. Cardiol.* 49.15 (2007), pp. 1651–1659.

[37] D. Liu, K. Hu, S. Störk, et al. "Predictive value of assessing diastolic strain rate on survival in cardiac amyloidosis patients with preserved ejection fraction". In: *PLoS One* 9.12 (2014), e115910.

[38] R. I. Williams, N. Payne, T. Phillips, J. D'hooge, and A. G. Fraser. "Strain rate imaging after dynamic stress provides objective evidence of persistent regional myocardial dysfunction in ischaemic myocardium: regional stunning identified?" In: *Heart* 91.2 (2005), pp. 152–60.

Appendix

Preprocessing

Velocity and deformation traces acquired during exercise had fewer samples than those acquired at rest (120 $\pm$ 31 vs 186 $\pm$ 45, P $<$ .0001), because the R-R intervals were shorter, so interpolation was performed independently for the rest and exercise data. Each of the six mechanical phases of the cardiac cycle [1] was interpolated separately, to have the same number of data points as the average number for that phase in all "healthy control" subjects.

Statistical Modeling for STRE Representation Based on PCA

The STRE representation of LV function of each subject consisted of 5,958 samples, with 3,744 samples from the resting curves (18 segmental curves each with 208 temporal instances) and 2,214 samples from the exercise curves (18 segmental curves each with 123 temporal instances). The rationale of using PCA for learning the STRE representations of the subjects was its capacity to model data of high dimensionality (i.e., with a large number of input parameters) compactly [2]. This feature allowed us to significantly reduce the initial dimensionality of the STRE data in order to build a classification system with our database consisting of 100 subjects.

Quality Control Indexes to Detect Artifactual Curves

We considered the following three indexes to detect artifactual and/or physiologically implausible segmental curves (equivalent to 9% of the resting data and 7% of the exercise data):

1. positive strain amplitude throughout most of the cardiac cycle (defined as $> 66\%$ of the RR interval);

2. strain amplitude that is more negative than -50% at any time during the cardiac cycle; and

3. strain rate amplitude that is more negative than -4/sec during early diastole (when longitudinal strain rate would normally be positive); the traces of all subjects are shown superimposed in Supplemental Figure 1, with a dashed horizontal line demonstrating the cut point of -4/sec to illustrate how its application will remove the extreme outlying traces.

These indexes were applied to the segmental curves acquired at rest and during exercise, and every curve that had any of the aforementioned index conditions was removed from the data set. If a strain trace met either or both conditions 1 and 2, then its corresponding velocity and strain rate traces were also removed. Likewise, if a strain rate trace met the third condition, its corresponding velocity and strain traces were omitted. The removed segmental traces were then approximated through KNNimpute and replaced in the data set.
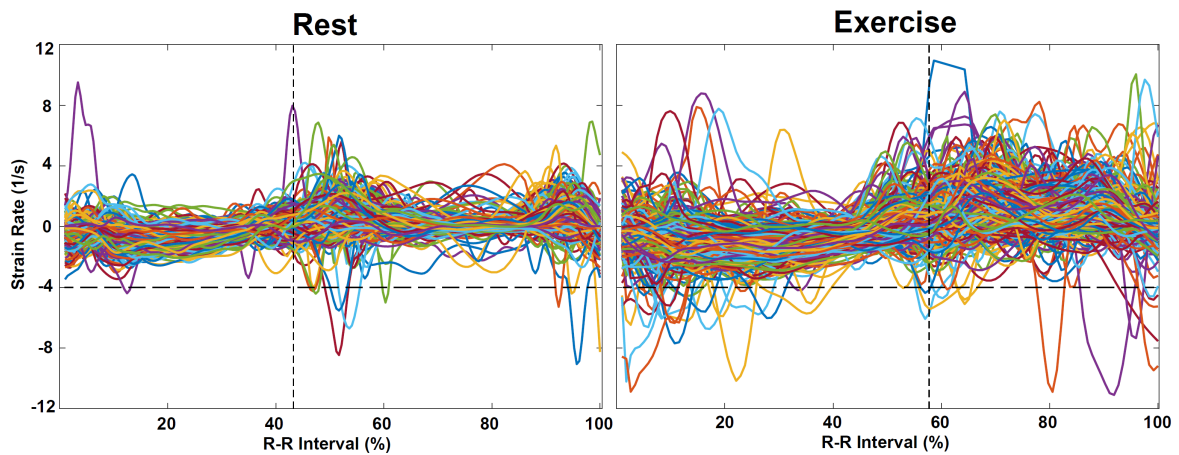
Figure 1: Superimposed segmental strain rate curves acquired at rest and during exercise. The dashed horizontal line shows the cutoff point for excluding the segmental curves with amplitude more negative than -4/sec during early diastole. End-systole is shown with a dashed vertical line.

Finding the Optimal Parameters of the Machine-Learning Framework

Supplemental Figure 2 illustrates an overview of the process of training and testing the machine-learning algorithm in 100 rounds. In each round, 16 unique subjects from each of the four groups were randomly selected to form the training set, and the remaining subjects formed the testing set. In this way, it is ensured that the learning system was evaluated on a set of unseen subjects (i.e., the testing set). The training data were used to build the PCA model and to compute an eigenvector matrix (i.e., matrix of PCs). Both the training and testing data were then projected onto the space spanned by the PCs and the obtained values (i.e., features) were used for training and testing a DWKNN classifier. We examined different number of features by preserving a variety of PCs, and for each DWKNN classifier, different KNNs were tested. After running the machine-learning algorithm 100 times, a combination of the number of PCs and KNNs that led to the best average testing result was selected.

The same strategy was used for selecting the optimal number of KNNs when the conventional features, extracted from the deformation curves, and baseline echocardiographic measurements were used for the classification task. More detailed explanations of the process of parameter selection and evaluation of a machine-learning algorithm are given in chapter 7 of The Elements of Statistical Learning [3].

Discriminating Symptomatic from Asymptomatic Subjects

The computed H values for the STRE and conventional features are presented in Supplemental Table 1. They are consistent with the results of the ROC analyses, in which the strain rate and velocity STRE features yielded the highest H values (0.508 and 0.353, respectively), and the conventional measurements of the velocity curves had higher H values than the other standard echocardiographic measurements. It is of note that H takes
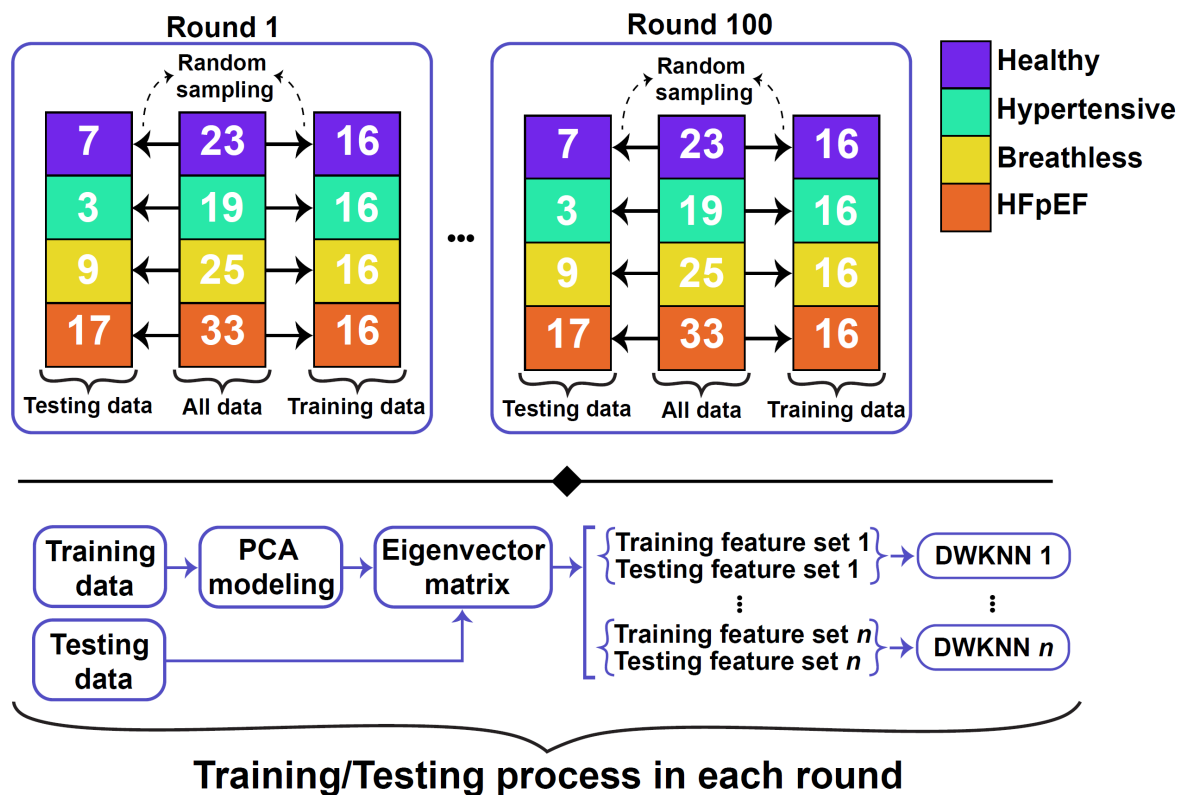
Figure 2: Process of finding optimal parameters of the machine-learning algorithm and evaluating its performance. In each round of the machine learning analysis, 64 unique subjects were randomly selected from the four groups (16 subjects per group) as the training data, and the remaining (36 subject) were used as testing. After building the PCA model and computing an eigenvector matrix, the training and testing data are projected onto the space spanned by the eigenvector matrix to obtain the feature vectors required for building a DWKNN classifier. After performing the above process in 100 rounds, a combination of the number of PCs and KNNs that resulted in the best average testing outcome was selected.

values between 0 (for worst classification performance) and 1 (for perfect classification), where higher performance corresponds to larger values [4].

Independent Statistical Modeling of the Rest and Exercise Data

We conducted an experiment to demonstrate if modeling the combined rest-exercise data (i.e., STRE) leads to better results than modeling the rest and exercise data independently. We first concatenated 18 segmental curves of each subject acquired at rest to form a spatiotemporal-rest (rest) representation of LV function. Similarly, a spatiotemporal-exercise (exercise) representation of LV function was formed by concatenating 18 segmental curves acquired during submaximal exercise. Independent PCA models were then built with the rest and exercise data to learn their main patterns. This resulted in six different sets of learned PCA parameters for the velocity, strain, and strain rate data and six distinct DWKNN classifiers that were constructed with these parameter sets for
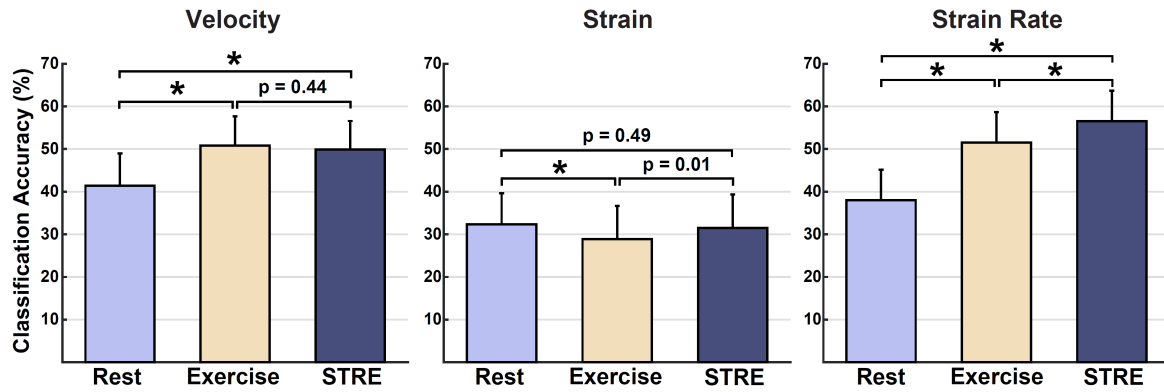
Figure 3: Diagnostic accuracy of spatiotemporal patterns of the rest, exercise, and combined rest-exercise data. The best classification accuracy rates (mean ± SD) obtained by our proposed machine learning method on the basis of modeling of the echocardiographic variables (velocity, strain, and strain rate) acquired either at rest or during exercise. The results are compared with machine learning of combined rest-exercise (STRE) modeling (∗ P < .0001).

solving the four-class classification problem. We repeated this procedure 100 times with the same splits used for evaluating the STRE modeling.

Supplemental Figure 3 shows the best average classification outcomes as well as the results of statistical comparison of the rest and exercise parameter sets with the STRE features. Although the strain rate STRE features could significantly improve the results obtained by the rest and exercise features, STRE modeling of the strain data did not yield better results than rest and exercise modeling. For the velocity data, STRE only outperformed rest, and its result was comparable to that of exercise. This observation again highlights the diagnostic value of the strain rate variable acquired at rest and during exercise.

Prediction of Exercise Capacity

The classification results obtained by applying thresholds equal to 350, 400, and 450 m to the 6-min walk test results are listed in Supplemental Table 2, Supplemental Table 3, Supplemental Table 4 for the velocity, strain, and strain rate curves, respectively. The tables show the AUCs, H values, and accuracy, sensitivity, and specificity of the classifiers built with the rest, exercise, and STRE data. For all examined thresholds, the STRE features of the strain rate curves yielded the highest results. Consistent with the results of the ROC analyses, the largest H values were obtained with the STRE features of the strain rate curves.

Influence of Frame Rate

The 11 subjects with low–frame rate acquisitions (mean, 130 ± 24 Hz) belonged to the hypertensive, breathless, and HFpEF groups. Before temporal alignment, their traces included 108 ± 22 sampling points at rest and 70 ± 11 during exercise; after alignment, the points were increased, as indicated in the methods, to 208 and 123. The cut point
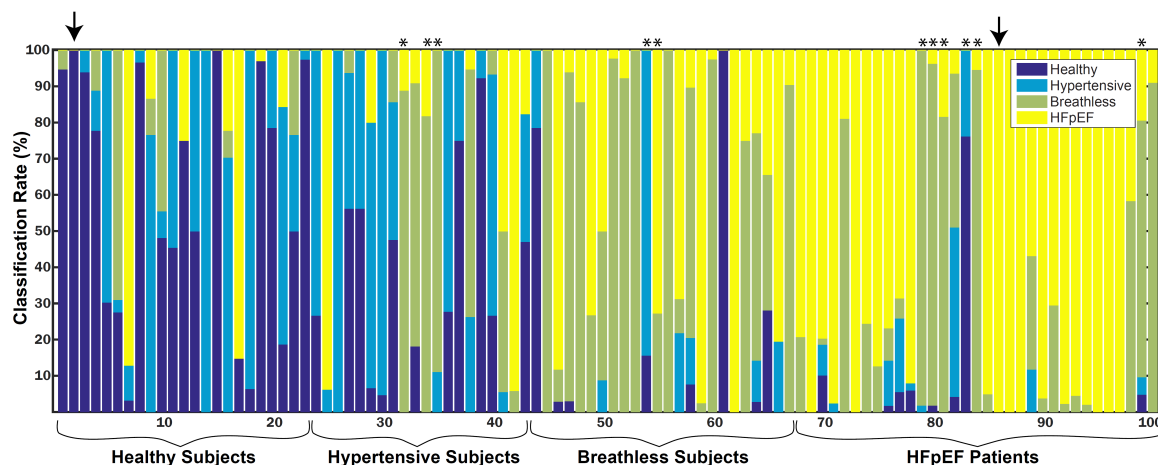
Figure 4: Stacked bar graph of the results obtained on repeated analyses (100 times) using strain rate STRE demonstrating the classification results for each subject. Eleven subjects with low acquisition frame rates are identified by the asterisks. The arrows show two subjects that were consistently classified by our machine-learning algorithm, and their concatenated segmental curves are displayed in Figure 7 in the main text.

for selecting low frame rates was set at 150 Hz for curves acquired either at rest or during exercise. To study the effect of low frame rate on the performance of the machine learning, a stacked bar graph was generated to display the classification results obtained with the strain rate STRE features, as the best feature set, in each subject, when the evaluation was repeated 100 times (Supplemental Figure 4).

From the total of 100 subjects, 39 were misclassified $> 70\%$ of the times that they were used in the test set. This group included all 11 subjects with low acquisition frame rates (i.e., 28% of the misclassified subjects); they are identified by the asterisks above the columns in Supplemental Figure 4. Six of these subjects had HFpEF, according to the clinical diagnosis; five were usually categorized as breathless subjects and one was categorized either as healthy or hypertensive. One breathless subject was classified as having HFpEF and another as healthy. Three hypertensive subjects were classified as breathless control subjects. These findings emphasize the importance of acquiring deformation curves at frame rates > 150 Hz, in order to identify dyssynchrony in patients with HFpEF.

To examine whether low acquisition frame rate affects the performance of machine learning, the PCA modeling of segmental strain rate (STRE features) and the automatic classification were repeated for the four-class case, after excluding the 11 subjects with low frame rates. This resulted in classification accuracy of $63 \pm 7\%$, instead of $57 \pm 7\%$ if the low frame rate subjects were included (P $<$ .0001). Thus, the diagnostic power of myocardial deformation imaging is related to the frame rate at which data are acquired. High temporal resolution is crucial for resolving the characteristics of short-lived events, especially during stress testing.

Physiological Interpretation of the Modeled STRE Patterns Using PCA

The first two PCs of the velocity, strain, and strain rate curves are visualized in Supplemental Figure 5, Supplemental Figure 6, Supplemental Figure 7. These PCs were more important than the others given that the PCs were sorted on the basis of their contributions in modeling data variations. Because the STRE data were generated by concatenating all the segmental curves from each subject acquired at rest and during exercise, each PC is broken into the rest and exercise phases and for each phase, the concatenated curves from the basal and mid segments are illustrated separately to enable easier reading of the PCs.

References

[1] J. D'hooge, B. Bijnens, J. Thoen, F. Van de Werf, G. Sutherland, and P. Suetens. "Echocardiographic strain and strain-rate imaging: a new tool to study regional myocardial function". In: *IEEE Trans. Med. Imaging* 21.9 (2002), pp. 1022–1030.

[2] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

[4] D. J. Hand. "Measuring classifier performance: a coherent alternative to the area under the ROC curve". In: *Mach. Learn.* 77.1 (2009), pp. 103–123.
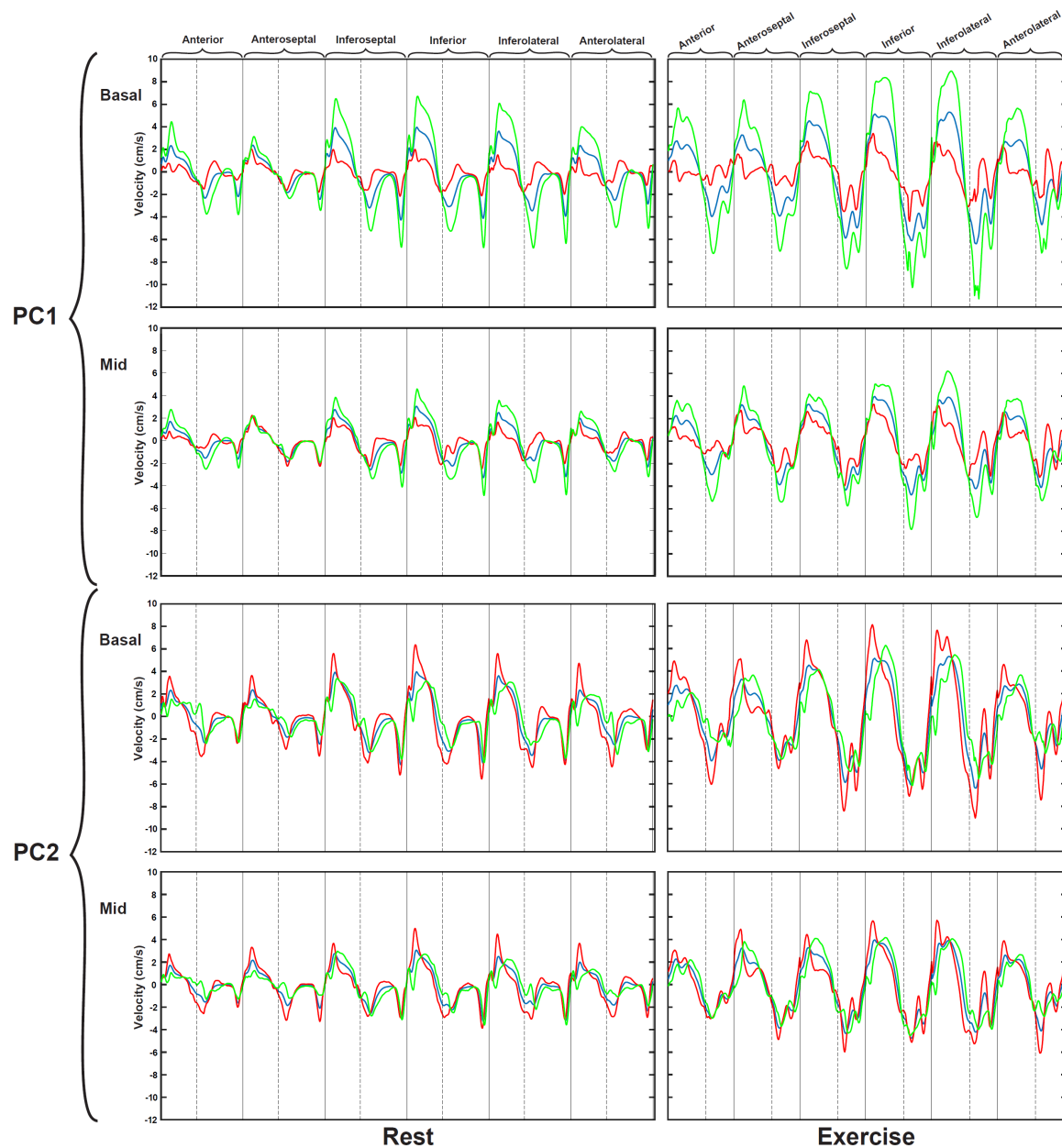
Figure 5: Mathematically reconstructed curves demonstrating the variations in the velocity curves modeled by PCA. Directions of variation of the first and second most important STRE patterns (i.e., PCs) of the velocity curves. The red and green curves are the results of adding and subtracting $\pm 2\sqrt{\lambda_i}$ of PC_i to the mean velocity curve, which is shown in blue. λ_i is the eigenvalue of PC_i, which demonstrates the amount of data variation modeled by the ith PC.

Figure 6: Mathematically reconstructed curves demonstrating the variations in the strain curves modeled by PCA. Directions of variation of the first and second most important STRE patterns (i.e., PCs) of the strain curves. The red and green curves are the results of adding and subtracting $\pm 2\sqrt{\lambda_i}$ of PC_i to the mean strain curve, which is shown in blue. λ_i is the eigenvalue of PC_i, which demonstrates the amount of data variation modeled by the ith PC.
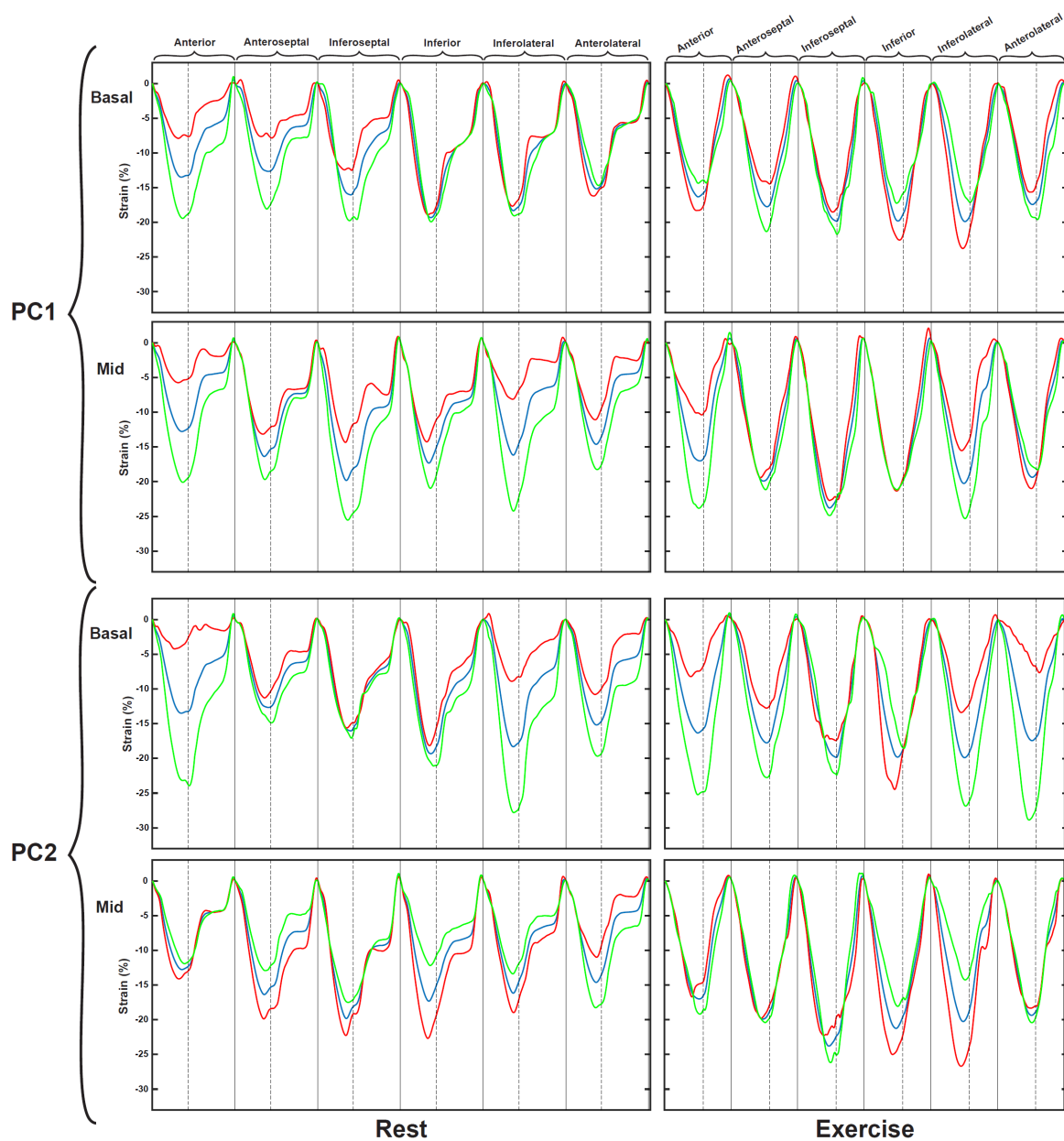
Figure 7: Mathematically reconstructed curves demonstrating the variations in the strain rate curves modeled by PCA. Directions of variation of the first and second most important STRE patterns (i.e., PCs) of the strain rate curves. The red and green curves are the results of adding and subtracting $\pm 2\sqrt{\lambda_i}$ of PC_i to the mean strain rate curve, which is shown in blue. λ_i is the eigenvalue of PC_i, which demonstrates the amount of data variation modeled by the ith PC.
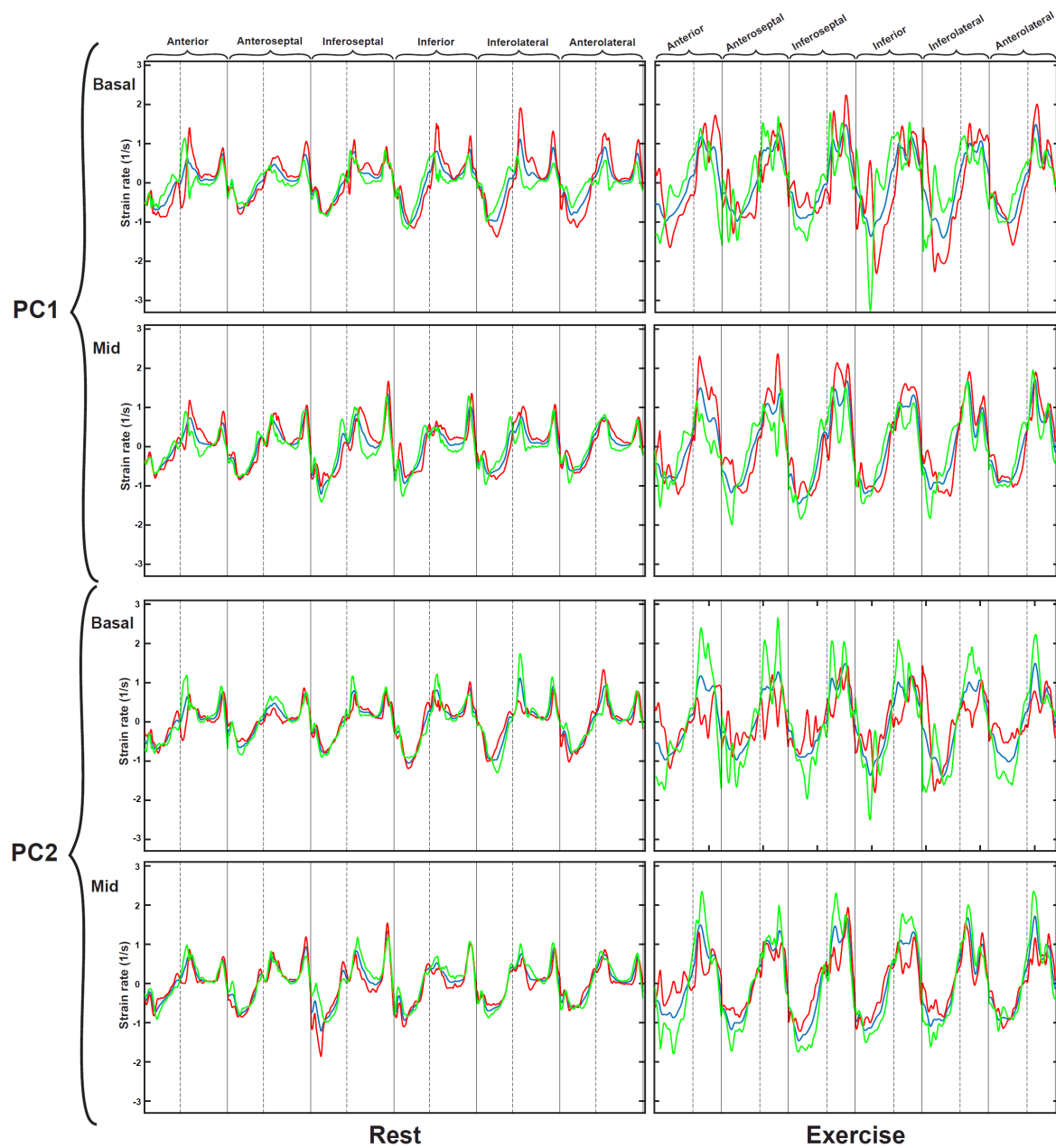
Table S1: H **measures of the classification outcomes of the 2-class case obtained with the STRE and with the conventional features of the velocity, strain and strain rate data**

	H measure
Velocity	
STRE	0.353
Peak systole	0.196
Time to peak systole	0.17
SD of time to peak systole	0.048
Peak early diastole	0.187
All markers	0.193
Strain	
STRE	0.133
Peak systole	0.018
Peak early diastole	0.054
1/3 diastole	0.085
All markers	0.068
Strain rate	
STRE	0.508
Peak systole	0.011
Time to peak systole	0.052
Peak early diastole	0.106
Time to peak early diastole	0.087
Peak late diastole	0.05
All markers	0.08

Table S2: Classification performances in predicting exercise capacity of the subjects using velocity curves

	AUC	*H* measure	Accuracy (%)	Sensitivity (%)	Specificity (%)
6MWT threshold = 350					
Rest	0.61	0.039	57 ± 7	58 ± 23	57 ± 9
Exercise	0.63	0.062	61 ± 7	62 ± 21	61 ± 8
STRE	0.67	0.090	63 ± 7	63 ± 23	63 ± 8
6MWT threshold = 400					
Rest	0.70	0.164	66 ± 8	67 ± 15	66 ± 10
Exercise	0.76	0.284	68 ± 8	69 ± 18	68 ± 12
STRE	0.78	0.298	70 ± 6	71 ± 15	70 ± 9
6MWT threshold = 450					
Rest	0.62	0.061	61 ± 7	62 ± 10	60 ± 16
Exercise	0.73	0.176	68 ± 7	68 ± 10	68 ± 16
STRE	0.73	0.178	68 ± 7	69 ± 11	68 ± 16

Table S3: Classification performances in predicting exercise capacity of the subjects using strain curves

	AUC	*H* measure	Accuracy (%)	Sensitivity (%)	Specificity (%)
6MWT threshold = 350					
Rest	0.56	0.017	54 ± 8	55 ± 24	54 ± 10
Exercise	0.60	0.040	59 ± 7	60 ± 19	59 ± 8
STRE	0.58	0.030	57 ± 9	58 ± 25	57 ± 10
6MWT threshold = 400					
Rest	0.62	0.068	59 ± 8	60 ± 18	59 ± 11
Exercise	0.65	0.106	64 ± 7	65 ± 14	64 ± 10
STRE	0.66	0.105	62 ± 7	63 ± 16	62 ± 10
6MWT threshold = 450					
Rest	0.59	0.051	56 ± 8	56 ± 11	54 ± 18
Exercise	0.55	0.011	55 ± 7	55 ± 10	53 ± 16
STRE	0.59	0.033	56 ± 7	57 ± 11	55 ± 15

Table S4: Classification performances in predicting exercise capacity of the subjects using strain rate curves

	AUC	*H* measure	Accuracy (%)	Sensitivity (%)	Specificity (%)
6MWT threshold = 350					
Rest	0.58	0.026	56 ± 9	57 ± 25	56 ± 10
Exercise	0.74	0.165	69 ± 9	69 ± 22	69 ± 10
STRE	0.73	0.153	68 ± 8	69 ± 21	68 ± 9
6MWT threshold = 400					
Rest	0.71	0.178	67 ± 9	67 ± 13	66 ± 11
Exercise	0.83	0.402	74 ± 7	74 ± 13	74 ± 11
STRE	0.86	0.442	77 ± 6	78 ± 13	77 ± 10
6MWT threshold = 450					
Rest	0.72	0.152	67 ± 8	67 ± 10	65 ± 16
Exercise	0.73	0.175	67 ± 8	67 ± 10	65 ± 16
STRE	0.77	0.229	69 ± 7	69 ± 10	68 ± 16