SUPPLEMENTAL MATERIAL

Supplemental Methods

I. Convolutional neural network

In the convolutional neural network (CNN), each bottleneck block **(Supplemental Figure I panel A)** in the time-dimension convolutional layers consists of two branches: (a) two convolutional layers, each of which consists of a batch normalization[32], a ReLU non-linear activation, and a one-dimensional (1D) convolution operating on the time dimension, (b) a shortcut connection with a 1x1 convolution followed by a max-pooling layer, where the shortcut connection allows information to flow easily from previous layers and enables easier optimization of deep neural networks[33]. The outputs from these two branches were added together to form the output of a bottleneck block. The time-dimension convolutional layers consist of several stacked bottleneck blocks, followed by an average-pooling layer. Next, the channel-dimension convolutional layer consists of a 1D convolution operating on the channel dimension, a batch normalization, and a ReLU non-linearity. Finally, a fully connected layer with sigmoid activation is used to obtain the predicted probability of AF recurrence for the input signal window. To mitigate overfitting problems, dropout layers[34] were inserted between the bottleneck blocks as well as before the last fully connected layer. The probabilities from all 5-sec windows of the same patient were averaged to obtain the final predicted probability of AF recurrence for each patient. For EGM-based CNN, model hyperparameters (i.e., kernel size, output channel size, number of bottleneck blocks, dropout probability) were configured to be the same as that in Attia et al.[14]; for ECG-based CNN, the number of bottleneck blocks were reduced to 6 (versus 9) because of the reduced number of 5-sec ECG windows for model training. Details of the CNN architecture can be found in **Supplemental Figure I panels B–C**.

II. APPLE and CHA2DS2-VASc scores

APPLE scores[3] are calculated as follows: (a) add one point for age > 65 years, (b) add one point for persistent or longs-tanding persistent atrial fibrillation (AF), (c) add one point for chronic kidney disease (CKD), (d) add one point if left atria (LA) diameter ≥ 43mm, (e) add one point if left ventricular ejection fraction (LVEF) < 50%.

CHA2DS2-VASc scores[4] are calculated as follows: (a) add one point for congestive heart failure (CHF), (b) add one point for hypertension, (c) add two points for age ≥ 75 years, (d) add one point for diabetes mellitus (DM), (e) add two points for TIA or CVA, (f) add one point for coronary artery disease (CAD) or peripheral artery disease, (g) add one point for age between 65 and 74 years, (h) add one point for female.

III. QRS subtraction

QRS subtraction for EGM was performed as follows. First, the Q, R, and S points in the three surface ECG channels (I, II, III) were identified. Second, the pythagorean distance from these QRS points was calculated. Third, the average Q-R time distance and the average R-S time distance were calculated, and a buffer of -40 millisecond from the Q point and a buffer of +40 millisecond from the S point were used to find the width of the QRS segment for all the three ECG channels. Fourth, the average QRS of each basket channel was calculated as follows: (a) find the corresponding QRS signal in the basket channel using the QRS time points obtained in prior steps, (b) remove baseline offset from each QRS by subtracting the first point from the signal, and (c) calculate the average QRS in this channel. Fifth, a bandpass filter of 0.5–100 Hz was applied to the signal. Sixth, the QRS segment was removed from the basket channels by subtracting the average QRS of the basket signal for each QRS time point. Lastly, all EGM signals were resampled to 200 Hz.

IV. Model hyperparameter tuning

Hyperparameters for CNNs were configured to be the same as that in Attia et al.[14] (except for reducing the number of bottleneck blocks in ECG-based CNN; see section I in Supplemental Methods) and no hyperparameter tuning was involved. Hyperparameters for the CatBoost[17] classifier were tuned using a grid search using 5-fold cross-validation on the training folds, including (a) number of iterations within {1000, 2000, 5000}, (b) learning rate within {0.001, 0.01, 0.1}, (c) depth within {2, 5, 10}, (d) L2 leaf regularization within {1, 3, 5}.

Supplemental Results

I. Effects of missing value imputation

To investigate the effect of our imputation method (i.e., imputing missing clinical values using the most frequent values) on model performance, we re-trained models involving clinical features using two additional imputation methods: (a) imputing missing values using the mean values and (b) imputing missing values using the median values. **Supplemental Table II** shows the model performance of these two imputation methods and their comparisons to the original imputation method (most frequent). There is no statistically significant difference among these three imputation methods. Moreover, **Supplemental Table III** shows performance of the models when trained only on patients without any missing clinical features (n = 114). Similarly, there is no statistically significant difference between performance on the entire cohort and performance on the subset without missing values. These results suggest that imputing missing values with the most frequent ones does not introduce bias in the input data in our study.

II. Model performance on patients with paroxysmal AF versus patients with non-paroxysmal AF

Here, we evaluate model performance on two subgroups of patients: patients with paroxysmal AF (n = 66) and patients with non-paroxysmal AF (n = 90) (i.e., persistent and long-standing persistent). As shown in **Supplemental Table IV**, fusion models result in similar prediction performance between the entire cohort and these two subgroups of patients (no statistically significant difference), and outperform clinical feature-based models and EGM/ECG-based models.

III. Model performance on patients with cryoablation versus patients with radiofrequency ablation

Furthermore, we evaluate model performance on patients with cryoablation (n = 38) and patients with radiofrequency ablation (n = 111). As shown in **Supplemental Table V**, fusion models result in similar prediction performance between the entire cohort and these two subgroups, and outperform clinical feature-based models and EGM/ECG-based models. We note that model performance on patients with cryoablation has high variance, which is likely due to the small number of patients who received cryoablation in our cohort (n = 38 in total; 4-5 patients in each test fold on average).

IV. Effects of the use of post-ablation ECGs

Among 156 patients in our cohort, 49 patients do not have pre-ablation ECGs. We used their ECGs in sinus rhythm immediately after ablation for these 49 patients. To examine if the use of post-ablation ECGs affects model performance, we evaluate the models using ECGs on the subset of patients whose pre-ablation ECGs are available (n = 107). As shown in **Supplemental Table VI**, models perform comparably between the entire cohort and patients with pre-ablation ECGs (no statistically significant difference).

V. Evaluation of calibration of models

In addition to discriminative measures (e.g., AUROC, sensitivity, and specificity), we evaluate the calibration of the models using Brier score[19] and expected calibration error (ECE)[20]. As shown in Supplemental Table VII, we observe that our EGM and ECG-based CNNs and fusion models have lower Brier scores and ECEs than APPLE score, CHASDS2-VASc score, and the clinical feature-based classifier. This suggests that our models not only provide higher prediction performance, but are also better calibrated than the existing clinical scores.

Supplemental Tables

Supplemental Table I. List of clinical features used in this study and the number of missing values in each feature.

Clinical Features	**No. Missing Values**
Type of AF	1
Age	0
Sex	0
Height	0
Weight	0
Body Mass Index (BMI)	0
Prior AF Ablation	0
Congestive Heart Failure (CHF)	1
Left Ventricular Ejection Fraction (LVEF)	11
Hypertension (HTN)	0
Hyperlipidemia (HLD)	0
TIA/CVA	1
Coronary Artery Disease (CAD)	0
History of Myocardial Infarction (MI)	0

Percutaneous Coronary Intervention (PCI)	0
Valvular Disease	0
History of Ventricular Tachycardia (VT)	0
Congenital Heart Disease	0
Asthma	0
Obstructive Sleep Apnea (OSA)	1
Diabetes Mellitus (DM)	3
Chronic Kidney Disease (CKD)	5
Left Atrial Surface Area From CT	5
Left Atrial Volume From CT	5
Left Atrial Volume From ECHO	24
Left Atrial Diameter > 42mm	1
Left Atrial Sphericity Index	6

Supplemental Table II. Comparison of different imputation methods on models using clinical features. Values are mean ± standard deviation across 10-folds.

*p-values compare between most frequent imputation and mean/median imputation. p-values are computed using DeLong's nonparametric test[35], where test fold patients' predictions in 10 folds are aggregated together prior to the test.

	Most Frequent Imputation (Original)	**Mean Imputation**		**Median Imputation**	
	AUROC	**AUROC**	**p-value***	**AUROC**	**p-value***
APPLE Score	0.644 ± 0.129	0.644 ± 0.129	1.000	0.644 ± 0.129	1.000
CHA2DS2-VASc Score	0.650 ± 0.133	0.650 ± 0.133	1.000	0.650 ± 0.133	1.000
Clinical Feature	0.755 ± 0.093	0.743 ± 0.108	0.385	0.770 ± 0.085	0.257
Fusion of EGM & Clinical Data	0.788 ± 0.110	0.781 ± 0.125	0.743	0.786 ± 0.113	0.620
Fusion of ECG & Clinical Data	0.836 ± 0.063	0.837 ± 0.050	0.995	0.840 ± 0.053	0.657
Fusion of EGM, ECG & Clinical Feature	0.859 ± 0.082	0.839 ± 0.093	0.745	0.854 ± 0.099	0.531

Supplemental Table III. Performance of models using clinical features on patients without any missing values. Values are mean ± standard deviation across 10-folds.

*p-values are computed using DeLong's nonparametric test[35], where test patients' predictions in 10 folds are aggregated together prior to the test.

	All patients (n = 156) AUROC	**Patients without any missing values (n = 114) AUROC**	**p-value***
APPLE Score	0.644 ± 0.129	0.654 ± 0.168	0.931
CHA2DS2-VASc Score	0.650 ± 0.133	0.691 ± 0.132	0.726
Clinical Feature	0.755 ± 0.093	0.733 ± 0.170	0.852
Fusion of EGM & Clinical Data	0.788 ± 0.110	0.777 ± 0.122	0.872
Fusion of ECG & Clinical Data	0.836 ± 0.063	0.834 ± 0.116	0.942
Fusion of EGM, ECG & Clinical Feature	0.859 ± 0.082	0.849 ± 0.100	0.946

Supplemental Table IV. Model performance on patients with paroxysmal AF versus patients with non-paroxysmal AF (persistent and long-standing persistent). Values are mean ± standard deviation across 10-folds. Best mean results for each group of patients are highlighted in bold. *p-values compare between all patients and patients with paroxysmal/non-paroxysmal AF. p-values are computed using DeLong's nonparametric test[35], where test patients' predictions in 10 folds are aggregated together prior to the test.

	All Patients (n = 156)	**Patients with Paroxysmal AF (n = 66)**		**Patients with Non-Paroxysmal AF (n = 90)**	
	AUROC	**AUROC**	**p-value***	**AUROC**	**p-value***
APPLE Score	0.644 ± 0.129	0.584 ± 0.224	0.603	0.630 ± 0.117	0.847
CHA2DS2-VASc Score	0.650 ± 0.133	0.634 ± 0.191	0.794	0.654 ± 0.206	0.878
Clinical Feature	0.755 ± 0.093	0.811 ± 0.158	0.864	0.673 ± 0.166	0.697
EGM	0.731 ± 0.105	0.641 ± 0.316	< 0.001	0.680 ± 0.248	0.004
ECG	0.767 ± 0.122	0.859 ± 0.128	0.335	0.575 ± 0.196	0.090
Fusion of EGM & Clinical Data	0.788 ± 0.110	0.774 ± 0.219	0.635	0.753 ± 0.234	0.730
Fusion of ECG & Clinical Data	0.836 ± 0.063	0.868 ± 0.164	0.880	0.780 ± 0.134	0.896
Fusion of EGM & ECG	0.833 ± 0.084	0.712 ± 0.242	0.781	**0.895 ± 0.113**	0.851
Fusion of EGM, ECG & Clinical Feature	**0.859 ± 0.082**	**0.909 ± 0.147**	0.431	0.871 ± 0.147	0.429

Supplemental Table V. Model performance on patients with cryoablation versus patients with radiofrequency ablation. Values are mean ± standard deviation across 10-folds. Best mean results for each group of patients are highlighted in bold.

[*]p-values compare between all patients and patients with cryoablation/radiofrequency ablation. p-values are computed using DeLong's nonparametric test[35], where test patients' predictions in 10 folds are aggregated together prior to the test.

	All Patients	**Patients with Cryoablation (n = 38)**		**Patients with Radiofrequency Ablation (n = 111)**	
	AUROC	**AUROC**	**p-value[*]**	**AUROC**	**p-value[*]**
APPLE Score	0.644 ± 0.129	0.583 ± 0.167	0.677	0.632 ± 0.129	0.825
CHA2DS2-VASc Score	0.650 ± 0.133	0.724 ± 0.334	0.388	0.644 ± 0.177	0.594
Clinical Feature	0.755 ± 0.093	0.693 ± 0.215	0.960	0.787 ± 0.098	0.985
EGM	0.731 ± 0.105	0.438 ± 0.399	< 0.001	0.729 ± 0.120	< 0.001
ECG	0.767 ± 0.122	0.469 ± 0.328	0.983	0.800 ± 0.117	0.812
Fusion of EGM & Clinical Data	0.788 ± 0.110	0.854 ± 0.256	0.880	0.751 ± 0.167	0.942
Fusion of ECG & Clinical Data	0.836 ± 0.063	0.760 ± 0.307	0.455	**0.896 ± 0.084**	0.831
Fusion of EGM & ECG	0.833 ± 0.084	0.906 ± 0.174	0.390	0.839 ± 0.108	0.598
Fusion of EGM, ECG & Clinical Feature	**0.859 ± 0.082**	**1.000 ± 0.000**	0.671	0.861 ± 0.090	0.821

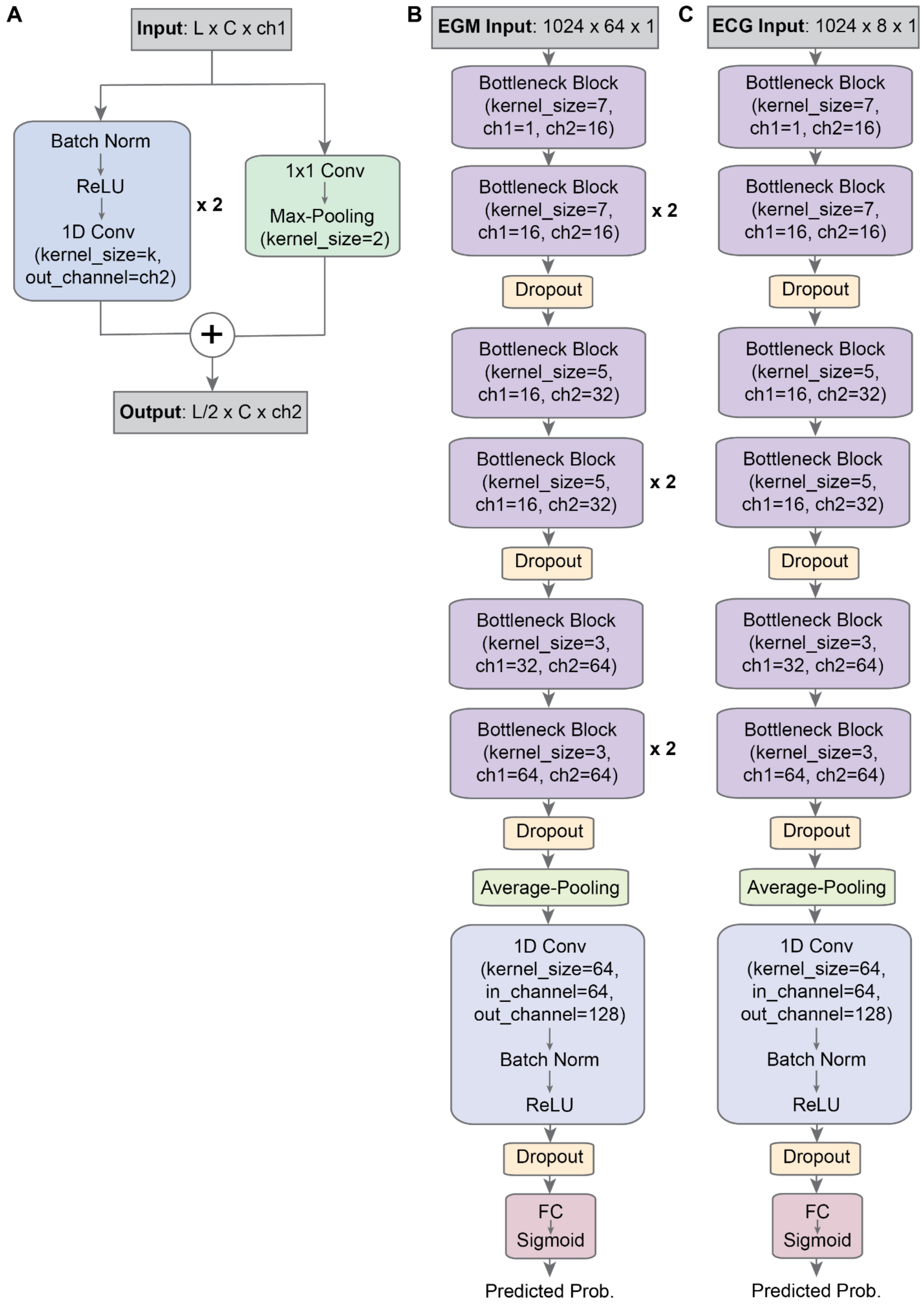
Supplemental Table VI. Performance of models using ECG data on patients whose prior-ablation ECGs are available. Values are mean ± standard deviation across 10-folds. *p-values are computed using DeLong's nonparametric test[35], where test patients' predictions in 10 folds are aggregated together prior to the test.

	All Patients (n = 156) **AUROC**	**Patients with** **Prior-Ablation ECG** **(n = 107)** **AUROC**	**p-value***
ECG	0.767 ± 0.122	0.677 ± 0.201	0.716
Fusion of ECG & Clinical Data	0.836 ± 0.063	0.810 ± 0.107	0.649
Fusion of EGM & ECG	0.833 ± 0.084	0.836 ± 0.117	0.918
Fusion of EGM, ECG & Clinical Data	0.859 ± 0.082	0.889 ± 0.125	0.723

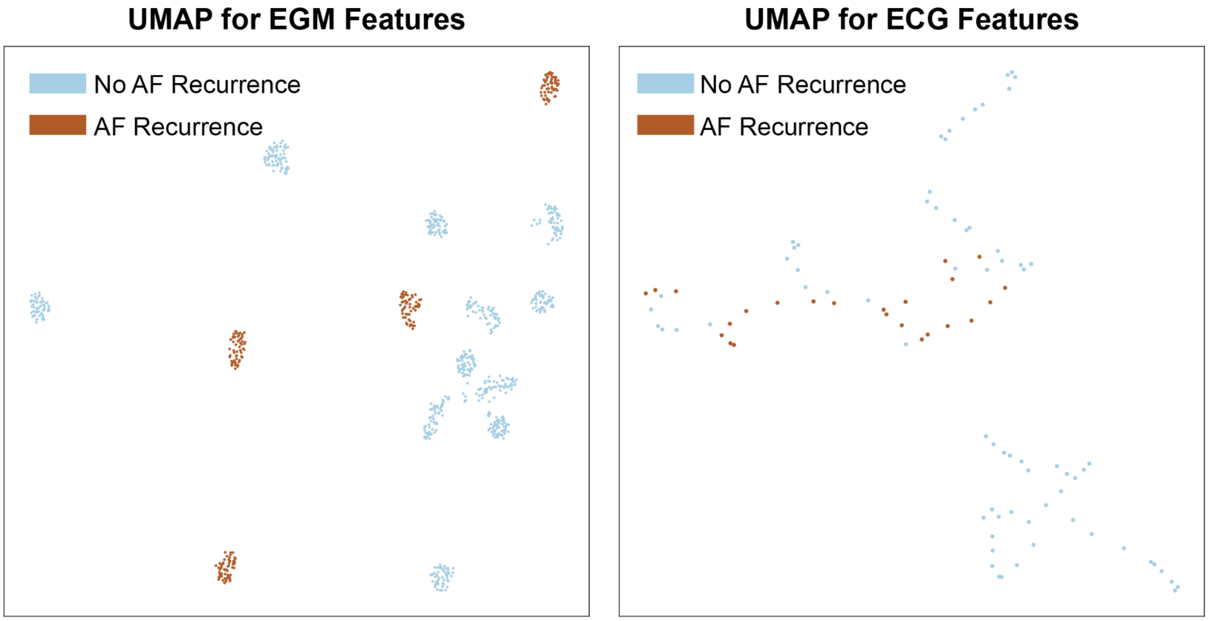
Supplemental Table VII. Brier score and expected calibration error (ECE) of baseline and our proposed models. Smaller Brier score and ECE indicate better calibrated models. Values are mean ± standard deviation across 10-folds.

	Brier Score	**Expected Calibration Error (ECE)**
APPLE Score	0.242 ± 0.013	0.199 ± 0.053
CHA2DS2-VASc Score	0.236 ± 0.022	0.193 ± 0.0363
Clinical Feature	0.231 ± 0.019	0.188 ± 0.043
EGM	0.198 ± 0.026	0.147 ± 0.034
ECG	0.199 ± 0.023	0.132 ± 0.048
Fusion of EGM & Clinical Data	0.210 ± 0.029	0.200 ± 0.059
Fusion of ECG & Clinical Data	0.201 ± 0.051	0.169 ± 0.053
Fusion of EGM & ECG	0.197 ± 0.037	0.186 ± 0.038
Fusion of EGM, ECG & Clinical Feature	0.206 ± 0.039	0.190 ± 0.061

Supplemental Figures and Figure Legends



Supplemental Figure I. (A) Illustration of a bottleneck block[33]. There are two branches. The first branch consists of two layers of (1) batch normalization[32], (2) ReLU nonlinearity activation, and (3) 1-dimensional convolution operated on the time dimension. The second branch consists of a shortcut connection with a 1x1 convolution and a max-pooling layer. The output from these two branches were added together to produce the output of the bottleneck block. The input has shape L x C x ch1, where L indicates the time dimension, C indicates the EGM/ECG lead/channel dimension, and ch1 indicates the input channel dimension. The resulting output has shape L/2 x C x ch2, where ch2 is the output channel dimension. **(B)–(C) CNN architecture for EGM and ECG**. The CNN consists of several bottleneck blocks operating on the time dimension with dropout layers, followed by an average-pooling layer and a final convolutional layer operating on the channel dimension. Finally, a fully connected (FC) layer with sigmoid activation is applied to produce the predicted probability of AF recurrence. Moreover, the input EGM or ECG matrices were padded to length 1024 in the time dimension with zeros to ensure that the max-pooling layers in the bottleneck blocks resulted in integer output lengths. We note that there are fewer ECG data (i.e., 5-sec windows) compared to EGM data, and thus the CNN architecture for ECG consists of fewer bottleneck blocks to reduce the number of model parameters.



Supplemental Figure II. Visualization of EGM (left) and ECG (right) features learned by the convolutional neural networks (CNN) using Uniform Manifold Approximation and Projection[23] (UMAP) in the test set of the median-performing folds (n=15 patients for EGM-based CNN and n=16 patients for ECG-based CNN). Specifically, the 128-dimensional EGM/ECG features are reduced to 2 dimensions using the UMAP dimensionality reduction algorithm and visualized here. Each dot represents features for one 5-second EGM/ECG window; blue dots represent patients without AF recurrence, while brown dots represent patients with AF recurrence. On the left panel, each cluster corresponds to one patient, suggesting that the same patient's features are more similar. On both panels, blue dots are further away from brown dots, suggesting that the models learned distinct features in patients with different outcomes.