uc3m | Universidad Carlos III de Madrid

Machine Learning in Healthcare

2025-2026

Final Project

"Forecasting Intracardiac Electrograms to Identify Arrhythmic Activity in Atrial Fibrillation"

Javier Fernández, Enica King, Virginia Fu, Sergio Madrid, Aleksander Nowak

Tutor

Gonzalo Ríos

Madrid, December 2025

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Atrial fibrillation (AF) is the most common type of cardiac arrhythmia, driven by regions of complex electrical activity inside the atria. **Intracardiac electrograms (EGMs)** recorded during **ablation procedures** allow us to access these dynamics. However, their interpretation is still challenging due to the high spatial and temporal variability of the signals.

The goal of this project is to apply **machine learning** and **deep learning-based forecasting** techniques to predict the temporal evolution of EGM signals. The working hypothesis is that signals that are **easier to predict** correspond to **passively activated regions**, while those that are **harder to forecast** reflect **active arrhythmic drivers**.

This approach may help identify relevant **ablation sites** for more personalised treatment strategies in atrial fibrillation.

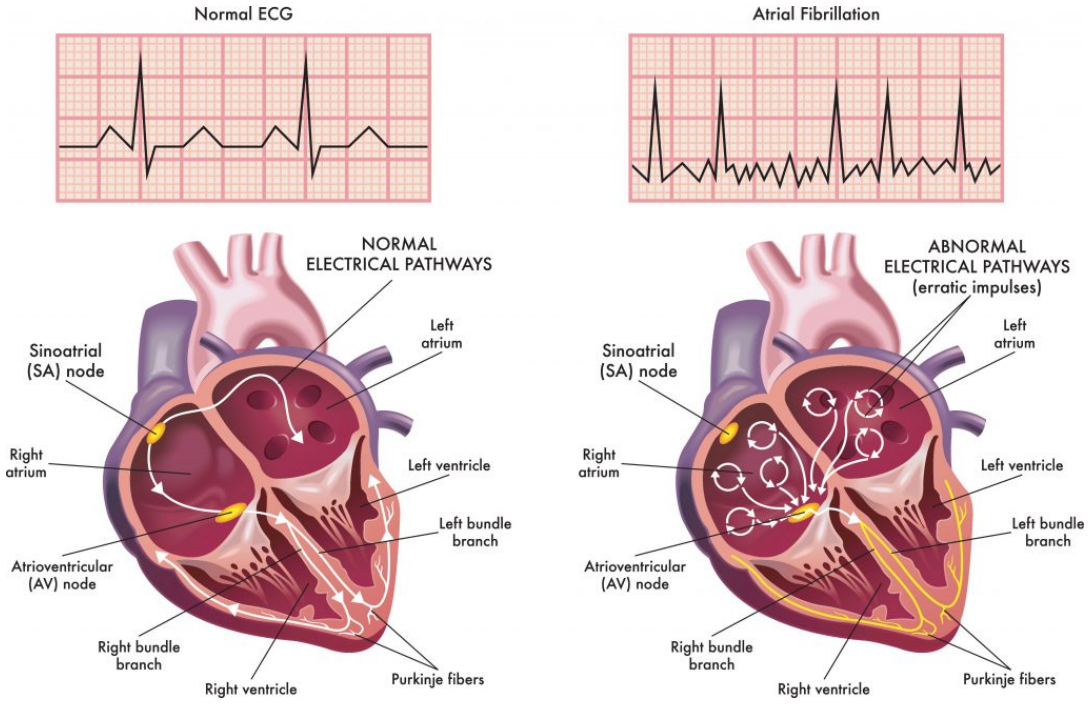

Figure 1: Comparison between normal sinus rhythm and atrial fibrillation.

1. INTRODUCTION AND PROBLEM DESCRIPTION

1.1. Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. This structure defines the mathematical foundation of our analysis:

- Ω is the sample space, representing all possible realizations of the heart signal.
- $\mathcal{F}$ is the σ-algebra, the collection of all measurable events.
- $\mathbb{P}$ is the probability measure, indicating the likelihood of specific signal patterns occurring.

We define the discrete-time intracardiac electrogram (EGM) signal as a real-valued stochastic process $(X_t)_{t\in\mathbb{Z}}$. This treats the EGM signal as a sequence of random variables indexed by time t. Since it is discrete-time ($t \in \mathbb{Z}$), it corresponds to the sampled data points (e.g., every 2 ms at 500 Hz).

The primary objective of this project was to develop a predictive model for these EGM signals. Our working hypothesis was that signals from healthy, passively activated tissue would be regular and predictable, whereas signals from active sources (arrhythmic drivers) would be chaotic and difficult to forecast. That is, a high prediction error could be used as a biomarker for identifying relevant ablation sites.

1.1.1. Initial Exploration and Baseline Models

Our initial approach focused on direct time-series forecasting using deep learning architectures. We established a pipeline that included:

- **Baselines:** Naive forecasting (repeating the last value) and Linear Regression.
- **Deep Learning Models:** Recurrent Neural Networks (RNNs), specifically Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks, as well as Sequence-to-Sequence (Seq2Seq) models with Attention mechanisms.

However, during this exploration, we encountered some problems. As observed in our initial experiments, the models frequently converged to predicting a "flat line" (the isoelectric line) rather than capturing the activation spikes. This happened because the Mean Squared Error (MSE) loss function penalizes large errors heavily. Since the activation

spikes are rare events in the temporal sequence, the model found it "cheap" to predict the mean value (zero) all the time instead of risking a large penalty by predicting a spike at the wrong time.

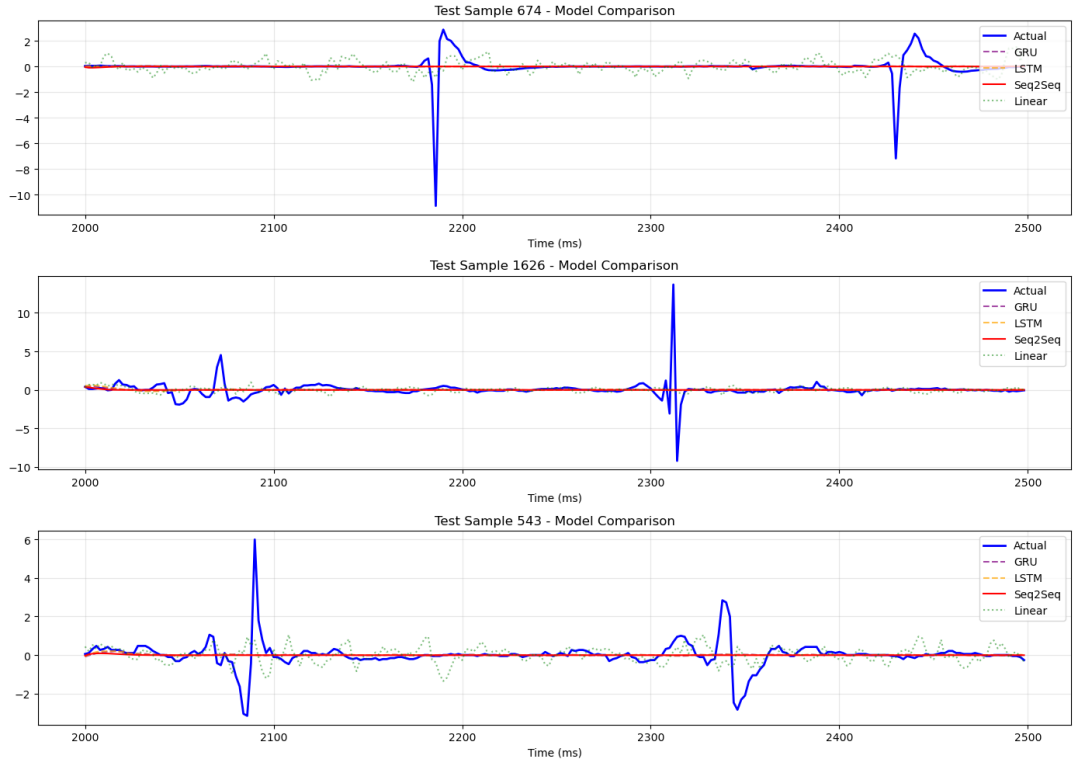

Figure 1.1: Comparison of deep learning model predictions (GRU, LSTM, Seq2Seq) vs. actual EGM signals.

The training process was stable for all deep learning models. As shown in the training curves, the loss decreased almost monotonically for the training and validation sets. Indeed, the models were successfully learning the underlying patterns of the data, even if they struggled with the sparsity of the spikes.

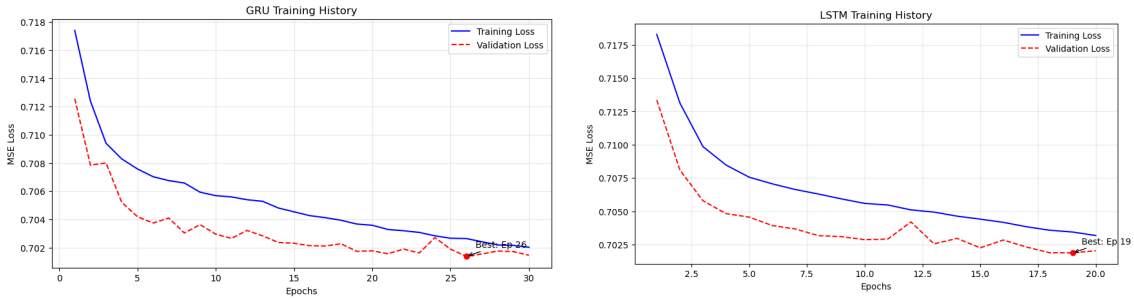

Figure 1.2: Training curves for GRU (left) and LSTM (right) models showing training and validation loss over epochs.

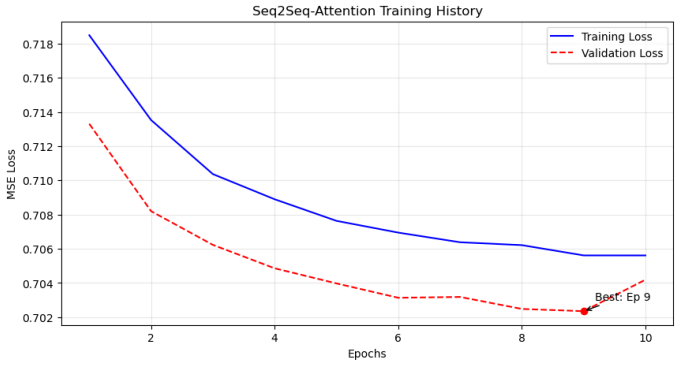

Figure 1.3: Training curve for the Seq2Seq model with Attention mechanism.

Additionally, there were stationarity issues, where signals would change rapidly from organized to fractionated in the same recording. There were also great amplitude differences between electrodes, which needed normalization. We also explored dimensionality reduction techniques, such as Principal Component Analysis (PCA), to simplify the signal complexity before forecasting.

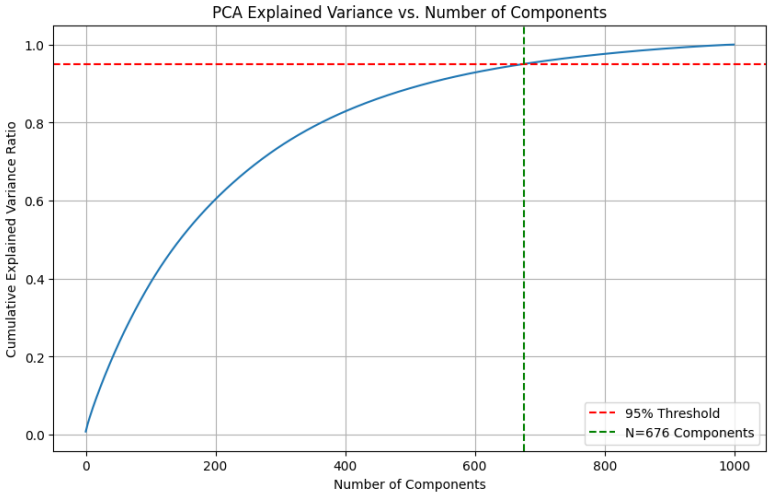

Figure 1.4: Principal Component Analysis (PCA) of signals showing the variance explained by each component.

When comparing model performance on PCA-transformed data versus the complete raw signals, PCA significantly improved the spike prediction metrics. With PCA, the Spike MSE ranged from 6.7–9.4 and Spike MAE from 2.4–2.9, whereas on the raw data these metrics were substantially worse (Spike MSE $\approx$ 16.5–16.8, Spike MAE $\approx$ 3.4–3.5). This suggests that dimensionality reduction helps the models focus on the most relevant signal components for predicting activations.

1.1.2. The Autoregressive Approach

Despite the improvements with PCA, we ultimately decided on an Autoregressive (AR) modeling strategy for several reasons:

1. **Instance-Specific Fitting**: Unlike neural networks that learn a global function across all patients, AR models fit coefficients independently for each signal. This avoids the need for a large training dataset and eliminates generalization issues across the different anatomies of patients.

2. **Prediction Error as a Biomarker**: The prediction error itself can be a potential biomarker. Arrhythmic tissue is expected to be harder to predict than healthy tissue. This approach favors more simple models in comparison to complex predictors.

3. **Interpretability**: AR coefficients have direct physical meaning related to tissue properties, whereas neural network weights are less explainable.

4. **Computational Efficiency**: AR models fit instantaneously via closed-form solutions (OLS), enabling real-time analysis of thousands of electrodes without GPU resources.

For this approach, we assume that $\mathbb{E}[X_t^2] < \infty$. This condition ensures that the second-order moments required for linear prediction exist.

The justification for this model comes from Wold's Decomposition Theorem [1]. Any zero-mean, non-deterministic stationary time series X_t can be expressed as an infinite moving average representation:

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \tag{1.1}$$

where $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$, and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a white noise process with zero mean and variance σ^2. Thus, the current value of the process can be expressed as a linear combination of current and past white noise terms.

The assumption of stationarity doesn't hold perfectly for EGM signals due to physiological variations. However, over short time intervals (e.g., 2.5 seconds), the stationarity assumption is a reasonable approximation that enables the use of autoregressive techniques.

The goal of this project is to take 2.5 seconds of EGM data at a sampling rate of 500 Hz, and use the first 2 seconds (Observation Window) to predict the next 0.5 seconds (Prediction Horizon).

2. DATA DESCRIPTION

In this section, we describe the EGM dataset used for analysis. The dataset has recordings from twelve patients, where each record corresponds to a specific map of the atria.

Let N_k be the number of observed processes (electrodes) for patient k, which varies depending on the location mapped during the procedure. Let T be the total time duration of the recording, which is constant across all recordings. For a given patient k, the dataset is represented as a matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{N_k \times T}$.

The data consists of EGM recordings sampled at a frequency of 500 Hz. Each recording has a duration of 2.5 seconds, corresponding to $T = 1250$ samples:

- Sampling Frequency (F_s): 500 Hz
- Sampling Interval (DT): 2 ms
- Observation Window (T_{OBS}): 1000 samples (2 seconds)
- Prediction Horizon (T_{PRED}): 250 samples (0.5 seconds)

We enforce the Weak Stationarity condition by centering the data to ensure $\mathbb{E}[X_t] \approx 0$. Specifically, we subtract the mean of the history window from the entire signal to avoid data leakage from the future.

The dataset is split into training and evaluation sets. However, for Autoregressive models, "training" is independent for each instance (fitting coefficients to the history of the specific signal being predicted). The split is for evaluating the generalization of hyperparameters (e.g., lag order p).

2.1. Signal Characteristics

Visual inspection of the EGM signals reveals three main patterns:

1. **Pattern A (Discrete, Organized Spikes)**: Characterized by sharp spikes separated by isoelectric baselines. These likely represent healthy or passively activated tissue.
2. **Pattern B (Complex Fractionated Atrial Electrograms)**: Characterized by continuous, low-amplitude activity without clear isoelectric lines. This can represent fibrotic tissue or active arrhythmic drivers.

3. **Artifacts**: Signals dominated by electrical interference or baseline drift, which are considered noise.

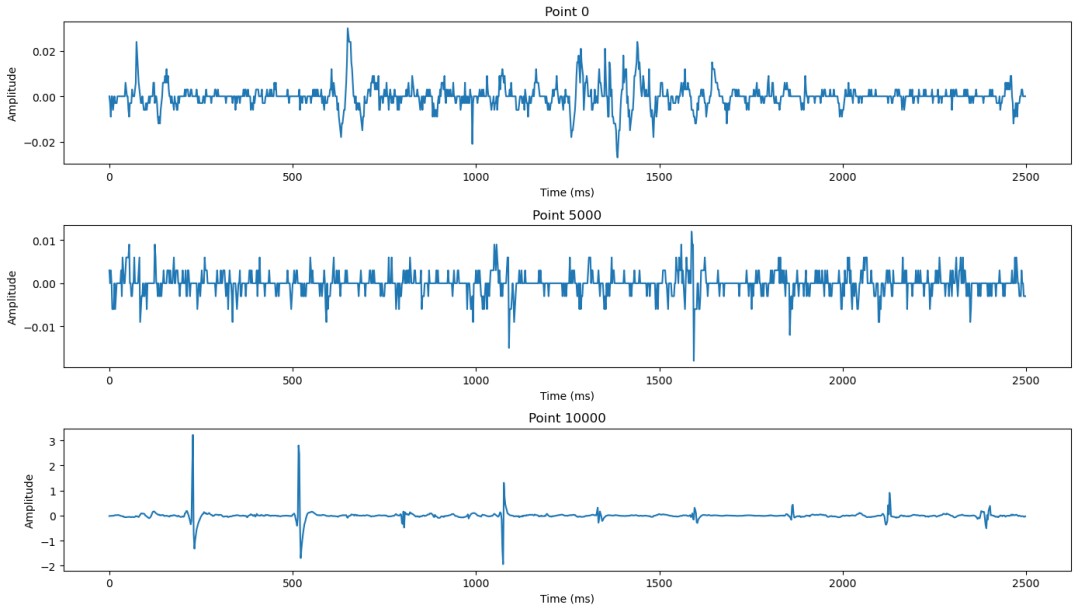

Figure 2.1: Sample EGM signals from the dataset showing the variability in waveforms across different electrodes.

3. PIPELINE AND METHODS PROPOSED

3.1. Autoregressive Process AR(p)

We approximate the Wold representation using an Autoregressive process of order p, denoted as AR(p). An AR(p) process describes X_t as a linear combination of its own past p values plus a stochastic error term:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t, \tag{3.1}$$

where c is a constant, $\phi_1, \ldots, \phi_p$ are the autoregressive coefficients, and ε_t is white noise. In training, we estimate the vector $\boldsymbol{\phi}$ using methods such as the Yule-Walker Equations or Least Squares Estimation (minimizing $\sum \varepsilon_t^2$).

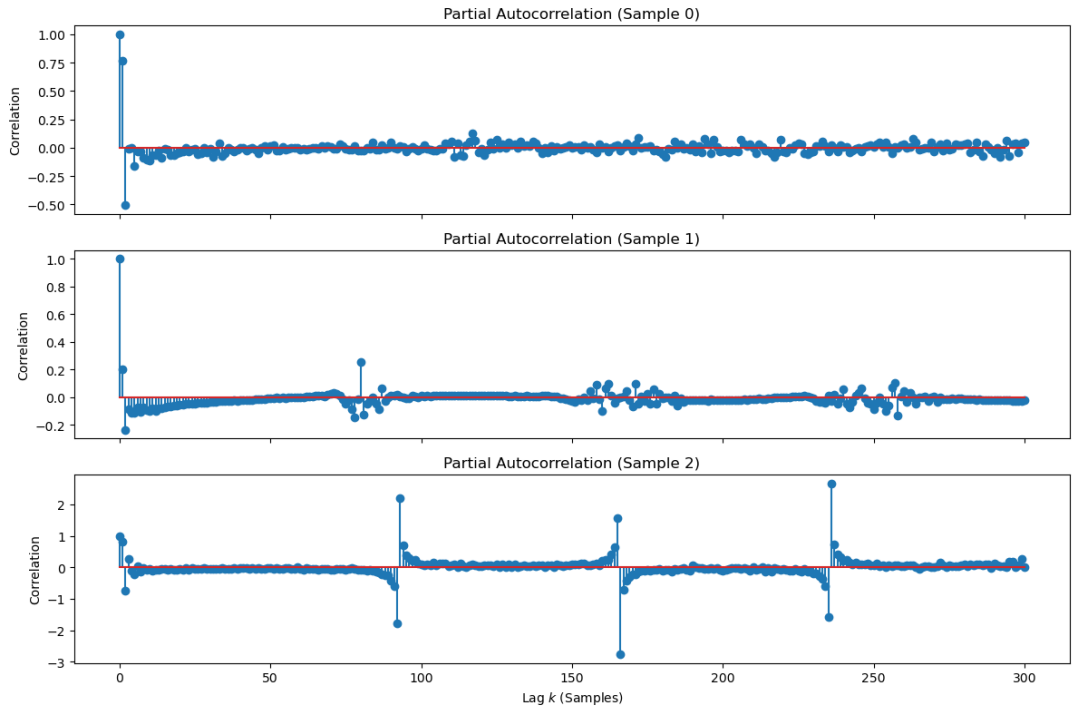

Figure 3.1: Autocorrelation function analysis used to determine the optimal lag order p for the AR model.

3.2. ARMA(p, q) and ARIMA(p, d, q)

To better capture shock dynamics, we extend the framework to the ARMA(p, q) model:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \tag{3.2}$$

where θ_j are the moving average coefficients. For non-stationary signals, we apply differencing d times, leading to the ARIMA(p, d, q) model:

$$(1 - L)^d X_t = c + \sum_{i=1}^{p} \phi_i (1 - L)^d X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \tag{3.3}$$

where $(1 - L)X_t = X_t - X_{t-1}$ is the differencing operator.

3.3. Time-Varying AR (TVAR)

Bio-signals often show non-stationarity. We generalize the AR(p) process to allow time-dependent coefficients:

$$X_t = c(t) + \sum_{i=1}^{p} \phi_i(t) X_{t-i} + \varepsilon_t. \tag{3.4}$$

The coefficients $\boldsymbol{\phi}_t$ are estimated recursively using Recursive Least Squares (RLS) via the Kalman Filter [2]. This allows the model to adapt to changing conditions (e.g., heart rate acceleration).

One of the main advantages of TVAR is its interpretability. We can visualize how the model adapts its parameters over the 2-second history window. If the coefficients remain flat, the signal was stationary. If they exhibit oscillations, the tissue properties were changing, which may indicate transitions between organized and fractionated activity.

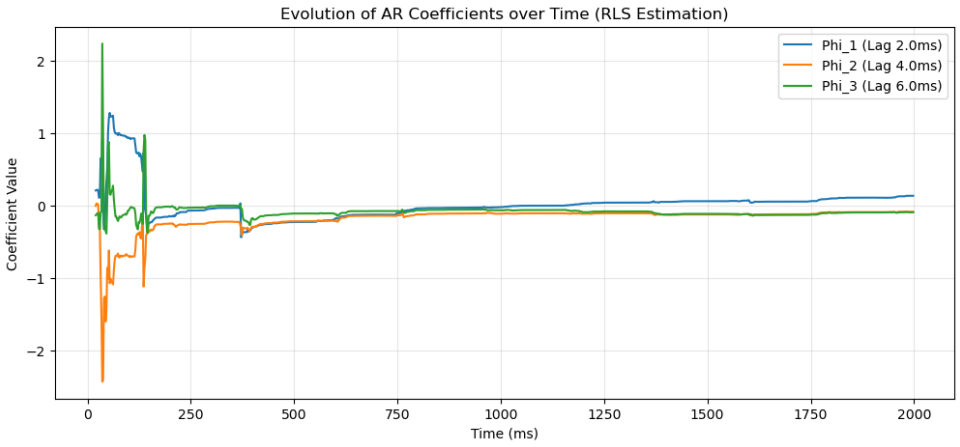

Figure 3.2: Time-varying AR coefficients $\phi_i(t)$ estimated via RLS.

3.4. Evaluation Metrics

Given the sparsity of EGMs, global MSE can be misleading. We introduce several metrics:

- **Global MSE**: Standard mean squared error over the entire prediction window.

- **Spike MSE**: MSE calculated only on high-amplitude regions (spikes), defined as samples exceeding the 95-th percentile of amplitude.
- **Normalized MSE (NMSE)**: Defined as NMSE = $\frac{\text{MSE}}{\text{Var}(y_{\text{target}})}$. This metric normalizes the error by the signal variance, providing an amplitude-invariant measure of prediction difficulty. An NMSE = 1 indicates the model performs no better than predicting the mean; NMSE > 1 indicates the signal is harder to predict than average.
- **Soft Dynamic Time Warping (SoftDTW) Loss**: We also explored SoftDTW [3], a differentiable variant of Dynamic Time Warping that measures similarity between time series while allowing for temporal shifts. This metric was promising, as it penalizes phase errors less harshly than MSE. However, it was computationally infeasible for our dataset due to its $O(N \cdot M)$ complexity per signal pair.

4. RESULTS AND DISCUSSION

4.1. Order Selection

We determined the optimal lag order p using the Akaike Information Criterion (AIC). Analysis of the dataset suggested an optimal order of $p \approx 10$ (covering 20 ms). This order was used for the AR models.

4.2. Stationarity Check

The Augmented Dickey-Fuller (ADF) test was performed on random samples. The results indicated that the majority of EGM signals in the dataset are stationary (p-value < 0.05). Therefore, for most cases, differencing ($d = 0$) is sufficient, favoring ARMA over ARIMA. However, we tested ARIMA with $d = 1$ for robustness.

4.3. Model Performance

We evaluated three model architectures:

1. **AR(p)**: Linear, static coefficients.
2. **ARIMA(p, d, q)**: Iterative, includes Moving Average terms.
3. **TVAR**: Time-varying coefficients estimated via RLS.

The results showed that while all models achieved reasonable Global MSE, they failed to accurately predict the "spikes" (activations) in the prediction horizon.

- **AR(p)**: Predictions tended to revert to the mean after approx. 100 ms.
- **ARIMA**: The addition of MA terms did not significantly improve spike prediction.
- **TVAR**: The RLS filter also struggled to predict future activations based solely on history, often producing a flat line or a very weak oscillation in the prediction window.

Visual inspection confirms that the models capture the periodicity to some extent but fail to reproduce the amplitude of the EGM spikes in the long term (> 100 ms). The

"Spike MSE" was significantly higher than the "Global MSE" for all models, making this limitation clear.

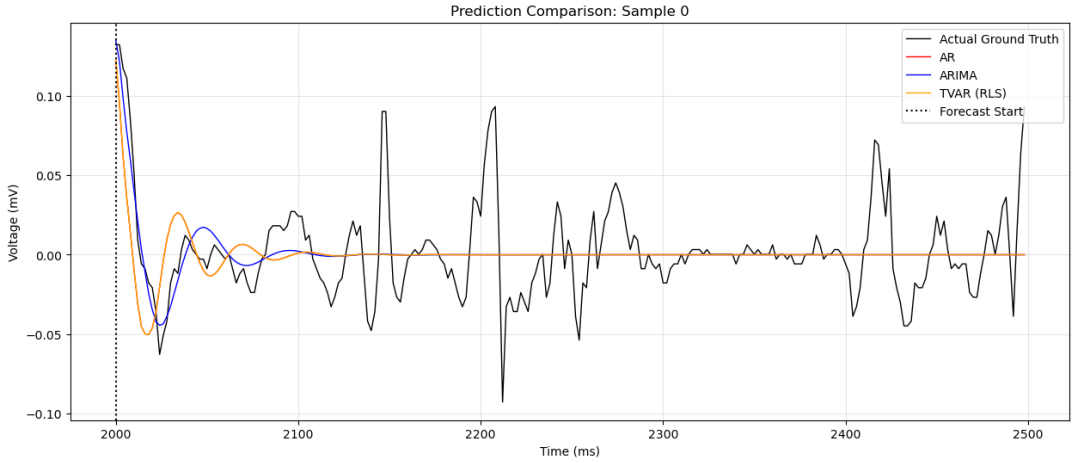

Figure 4.1: Comparison of AR model predictions vs. actual EGM signals.

4.4. Proposed Solution: Anomaly Detection via Prediction Error

Although the AR models struggle to produce accurate spike predictions. Our main insight is that **the prediction error itself becomes a biomarker for tissue complexity**.

We define the **Normalized Mean Squared Error (NMSE)** as:

$$\text{NMSE} = \frac{\text{MSE}}{\text{Var}(y_{\text{target}})} \tag{4.1}$$

where y_{target} is the ground truth signal in the prediction window. This normalization ensures that we rank electrodes by how chaotic the signal is, not by raw amplitude.

4.4.1. Hypothesis

- **Low NMSE** (< 1): Organized, predictable tissue. The AR model can track the signal reasonably well, suggesting passively activated or healthy regions.
- **High NMSE** (> 1): Complex, unpredictable tissue. The AR model fails, suggesting Complex Fractionated Atrial Electrograms (CFAEs) or drivers that are potential ablation targets.

4.4.2. Results of Anomaly Detection

We ran the AR model on the entire test set ($\approx 16,000$ electrodes across 12 patients) and computed NMSE for each electrode. The distribution revealed:

- The majority of electrodes cluster around NMSE ≈ 1, indicating that EGM signals are normally too stochastic for simple linear forecasting over a 0.5 second horizon.
- A clear **high-error tail** identifies **anomaly candidates**. These are signals significantly more unpredictable than the rest.
- A **low-error tail** (NMSE < 1) emphasizes organized, healthy tissue where rhythmic patterns allow for better predictions.

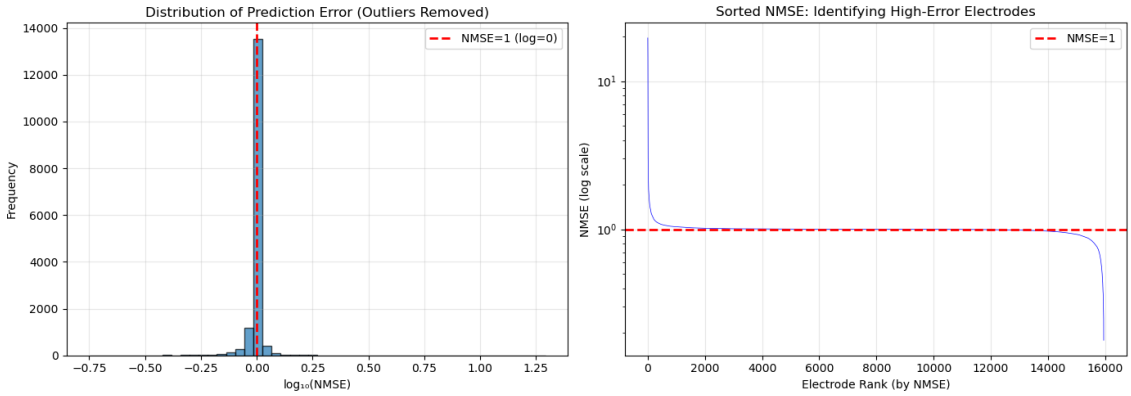

Figure 4.2: Distribution of Normalized MSE (NMSE) across all test electrodes in log-scale.

4.4.3. Visual Validation

Visual inspection of the top-ranked (high NMSE) signals confirmed that they exhibit characteristics of CFAEs: continuous, low-amplitude, fractionated activity without clear isoelectric baselines. Conversely, the bottom-ranked (low NMSE) signals showed discrete, organized spikes with clear baselines. These are consistent with healthy tissue.

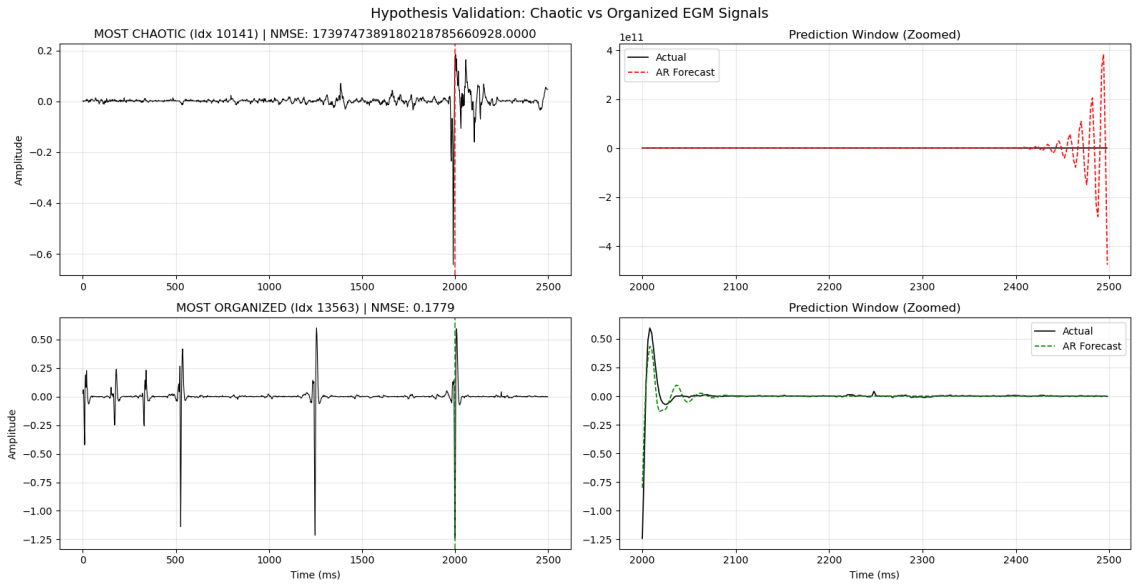

Figure 4.3: Comparison of high-NMSE vs. low-NMSE signals.

4.4.4. Candidate Ablation Sites

Based on the NMSE ranking, we identified the top 10% of electrodes by prediction error as **candidate ablation sites**. These regions, distributed across multiple patients and maps, represent areas where the AR model fails most severely and are therefore hypothesized to correspond to active arrhythmic drivers.

Table 4.1 presents a sample of the most unpredictable (potential arrhythmic drivers) and most predictable (healthy tissue) electrodes from the test set, after excluding three extreme outliers caused by numerical instability in the AR fitting process.

Table 4.1: Top 5 most unpredictable and most predictable electrodes by NMSE.

Patient Date	Map Name	Electrode	NMSE	MSE
Most Unpredictable (Potential Arrhythmic Drivers)				
2021_10_19	1-AI	0	19.55	0.0451
2021_11_17	1-AI BASAL FA	244	18.28	0.0024
2021_11_08	1-AI	445	18.04	0.0003
2021_11_08	1-AI	4517	17.61	0.1038
2021_10_18	1-1-1-FA	409	14.44	0.0001
Most Predictable (Healthy / Passively Activated Tissue)				
2021_11_08	1-AI	2567	0.178	0.0022
2021_10_19	1-AI	293	0.203	0.0001
2021_10_19	1-AI	3218	0.246	0.0001
2021_11_08	1-AI	2503	0.255	0.0002
2021_11_08	1-AI	2580	0.271	0.0001

We can observe that there is a meaningful difference in NMSE between the most unpredictable and most predictable electrodes: the highest NMSE (19.55) is nearly 110 times larger than the lowest (0.178). Second, the most predictable electrodes are concentrated in a single map (1-AI), suggesting that this region corresponds to healthy, passively activated tissue with regular activation patterns. Third, the candidate ablation sites are from four different patients and multiple maps, showing that the anomaly detection method generalizes across patients and is not biased toward a single recording session or anatomy.

5. CONCLUSIONS

In this study, we explored the use of Autoregressive (AR), ARIMA, and Time-Varying AR (TVAR) models for predicting intracardiac electrogram (EGM) signals and proposed an approach for identifying ablation targets.

Our analysis leads to the following conclusions:

1. **Short-term vs. Long-term Prediction**: AR-based models are effective for very short-term prediction (next few samples) but fail to capture the complex, non-linear dynamics required for a 0.5 second forecast. The models revert to predicting the mean (isoelectric line) after approximately 100 ms.
2. **Spike Prediction Limitation**: The models struggle to predict the sharp activations (spikes) that are clinically relevant. Increasing complexity from AR to ARIMA or TVAR provided little improvement.
3. **Prediction Error as a Biomarker**: We demonstrated that the **Normalized MSE (NMSE)** serves as a measure of signal "unpredictability." High NMSE values correlate with complex, fractionated activity (potential arrhythmic drivers), while low NMSE values correspond to organized, discrete spikes (healthy tissue).
4. **Clinical Application**: By ranking electrodes by NMSE, we identified candidate ablation sites, regions where the AR model fails most severely.

5.1. Limitations

- The test set is a subset of the full data; clinical validation would require correlation with actual ablation outcomes.
- The NMSE threshold for "chaotic" would need clinical calibration.
- Patient-specific factors (e.g., scar patterns) are not explicitly modeled.
- A small number of electrodes produced extreme NMSE values due to numerical instability in the AR fitting process. These were identified and excluded from the analysis.

5.2. Future Work

Future work should investigate:

- Correlation of high-NMSE regions with clinical ablation outcomes to validate the biomarker hypothesis.
- Spatial analysis to identify clusters of high-NMSE electrodes that may represent rotor cores.
- Patient-specific thresholds calibrated to individual anatomy and arrhythmia characteristics.

BIBLIOGRAPHY

[1] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Springer Series in Statistics, New York: Springer-Verlag, second ed., 1991. Section 5.7.

[2] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, p. 35, 1960.

[3] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," 2018.