

Hybrid Lagged Regression Modeling for Cardiac Anomaly Detection in ECG Time Series

¹Akinbo Yinka Rasaki; ²Timothy O Olatayo; ³Oseni Azeez Bolalale; ⁴Idowu Hammed Olabode; ⁵Adesanya Kehinde Kazeem;

^{1, 4 and 5} Ogun State Polytechnic of Health and Allied Sciences, Ilese-Ijebu

² Olabisi Onabanjo University (OOU), Ago-Iwoye

³Federal University of Agriculture and Technology, Okeho, Oyo State.

Abstract

Continuous cardiac monitoring generates large volumes of electrocardiogram (ECG) and related physiological time-series data that provide opportunities for the early detection of anomaly cardiac patterns. However, the temporal dependence, delayed physiological responses, multivariate interactions, non-stationarity, and nonlinear characteristics of cardiac signals present challenges for conventional statistical and anomaly-detection methods. This study developed a Hybrid Lagged Regression Model (HLRM) for modeling continuous cardiac time-series data and detecting cardiac anomalies. Secondary cardiac time-series data derived from the MIMIC-IV database through PhysioNet were used. The framework integrates autoregressive persistence, distributed lag effects, regularization, and mild nonlinear components within an interpretable statistical model. Its performance was evaluated against Autoregressive (AR), Distributed Lag (DLM), Vector Autoregressive (VAR), LASSO, and Elastic Net models. Stationarity was assessed using the Augmented Dickey-Fuller and KPSS tests, while predictive performance was evaluated using mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). The HLRM achieved an R^2 of 0.742 and an adjusted R^2 of 0.718 and produced the lowest prediction errors among the evaluated models, with $MSE = 0.024$, $RMSE = 0.155$, and $MAPE = 7.9\%$. For anomaly detection, the model achieved sensitivity = 0.89, specificity = 0.84, precision = 0.81, recall = 0.89, and area under the receiver operating characteristic curve (AUC) = 0.92. These findings demonstrate the potential of combining temporal persistence, distributed lag relationships, regularization, and nonlinear effects within a unified and interpretable framework for continuous cardiac anomaly detection. The HLRM provides a promising statistical approach for supporting early-warning applications in continuous cardiac monitoring.

Keywords: Cardiac anomaly detection, Electrocardiogram (ECG), Hybrid Lagged Regression Model, Regularization, Time-series analysis.

Date of Submission: 16-09-2026

Date of Acceptance: 29-09-2026

I. Introduction

Cardiovascular diseases (CVDs) remain a major global health burden and are associated with substantial mortality, morbidity, healthcare expenditure, and loss of productivity. Conditions such as coronary artery disease, cardiac arrhythmias, heart failure, and myocardial infarction can result in severe outcomes when anomalies are not identified and managed early. Consequently, timely identification of anomalous cardiac activity remains an important component of preventive and proactive cardiovascular care. Advances in wearable health technologies, bedside monitoring systems, and the Internet of Medical Things (IoMT) have increasingly enabled continuous collection of physiological measurements such as electrocardiogram (ECG) signals, heart rate, heart rate variability (HRV), blood pressure, and blood oxygen saturation (Dunn et al., 2018). These developments provide opportunities to move beyond episodic clinical assessment toward continuous monitoring capable of identifying subtle changes in cardiac function before severe clinical deterioration occurs.

Continuous cardiac data, however, present considerable analytical challenges because physiological measurements recorded over time cannot necessarily be treated as independent observations. Cardiac time series may exhibit autocorrelation, temporal variability, trends, circadian patterns, measurement noise, and delayed physiological responses (Box et al., 2015; Shaffer & Ginsberg, 2017; Rajendra Acharya et al., 2006). Heart rate and HRV also reflect complex interactions between sympathetic and parasympathetic activity, and changes in their temporal dynamics may contain physiologically relevant information that is not apparent from isolated measurements (Task Force, 1996; Malik & Camm, 2004). Consequently, conventional approaches based primarily on point-in-time observations or summary statistics may discard valuable temporal information contained in

continuous physiological signals (Clifford et al., 2006; Goldberger et al., 2002). Appropriate modeling of these temporal relationships is therefore important for identifying patterns that may precede anomalous cardiac states.

Time-series methods provide a useful statistical framework for characterizing such relationships. Autoregressive (AR) models relate the current value of a physiological signal to its previous observations and can therefore represent short-term temporal persistence. Lagged regression extends this principle by incorporating previous values of explanatory variables when modeling a current outcome, allowing delayed relationships among physiological processes to be quantified explicitly (Shumway & Stoffer, 2017). Distributed Lag Models (DLMs) can estimate the influence of explanatory variables over several time lags, while Vector Autoregressive (VAR) models permit dynamic interactions among multiple time-dependent physiological measurements (Hamilton, 1994; Lütkepohl, 2005). These methods are particularly attractive for biomedical applications because the estimated temporal relationships can be examined explicitly rather than being concealed within a complex predictive system.

Nevertheless, incorporating numerous candidate variables and their lagged values can substantially increase model dimensionality and introduce strong correlations among predictors. These conditions can produce unstable coefficient estimates and increase the risk of overfitting. Regularization techniques provide one approach to controlling these problems. The Least Absolute Shrinkage and Selection Operator (LASSO) applies coefficient shrinkage and variable selection, allowing less informative predictors to be excluded from the final model (Tibshirani, 1996). Elastic Net combines L1 and L2 regularization and is particularly useful where groups of predictors are highly correlated, as may occur among consecutive physiological lags (Zou & Hastie, 2005; Hastie et al., 2009). Regularization can therefore complement conventional time-series methods when multiple lagged predictors are considered simultaneously.

Machine-learning and deep-learning methods have also become increasingly important in automated ECG analysis and cardiac anomaly detection. Their ability to learn complex and nonlinear representations makes them attractive for analyzing physiological signals. However, improved predictive flexibility can be accompanied by reduced transparency. In clinically oriented applications, it may be important not only to predict whether an observation is anomaly but also to understand which physiological variables and temporal relationships contributed to that prediction. Concerns surrounding the interpretability of black-box predictive models have consequently encouraged the development of approaches that seek to combine predictive performance with transparent and explainable model structures (Rudin, 2019). This trade-off is particularly relevant to cardiac monitoring, where clinically meaningful temporal relationships may themselves provide valuable information.

Despite advances in cardiac time-series analysis, important methodological gaps remain. Previous research has frequently focused on discrete cardiac events, relatively short observation windows, or individual components of temporal behavior rather than integrating several temporal mechanisms within a unified framework (Cerqueira et al., 2020; Clifford et al., 2016). Univariate AR and ARIMA approaches may adequately represent short-term temporal dependence but can overlook multivariate interactions and distributed physiological lags. DLMs provide greater flexibility for delayed relationships but can become difficult to estimate when numerous correlated lagged variables are included. Similarly, although machine-learning and deep-learning methods can achieve strong predictive performance, limited transparency may restrict physiological interpretation in clinically oriented applications (Hannun et al., 2019; Rudin, 2019). The characterization and selection of informative temporal lags in continuous physiological streams also remain important methodological considerations.

These limitations suggest the need for an interpretable framework that combines complementary modeling strategies rather than addressing temporal persistence, distributed effects, predictor selection, and nonlinear relationships independently. In particular, the integration of autoregressive persistence, distributed lag effects, regularization, and mild nonlinear components within a unified statistical framework remains comparatively underexplored for continuous cardiac anomaly detection. Such integration could provide greater flexibility than conventional individual time-series models while maintaining a level of interpretability that is difficult to achieve with highly complex black-box approaches.

Accordingly, this study proposes a Hybrid Lagged Regression Model (HLRM) for modeling continuous cardiac time-series data and supporting cardiac anomaly detection. The HLRM integrates autoregressive components to represent temporal persistence, distributed lag terms to characterize delayed physiological relationships, regularization to control multicollinearity and select informative lagged predictors, and mild nonlinear components to accommodate departures from purely linear relationships. Cardiac time-series data derived from the MIMIC-IV database available through PhysioNet are used to evaluate the proposed framework against conventional Autoregressive (AR), Distributed Lag (DLM), Vector Autoregressive (VAR), LASSO, and Elastic Net models. The study specifically seeks to characterize temporal and lagged relationships in cardiac measurements, evaluate the predictive and anomaly-detection performance of the proposed HLRM, and determine its comparative performance against conventional and regularized time-series approaches. By integrating complementary sources of temporal information within an interpretable statistical framework, the study aims to contribute a flexible methodological approach for continuous cardiac monitoring and the development of data-driven early-warning systems.

II. Models Specification

2.1 Autoregressive Model

An Autoregressive (AR) model was fitted as a baseline approach for capturing temporal persistence in the cardiac response. An AR model of order p was specified as

$$Y_t = \alpha + \sum_{i=1}^p \phi_i Y_{t-i} + \varepsilon_t,$$

where Y_t denotes the cardiac response at time t , α is the intercept, ϕ_i is the autoregressive coefficient associated with lag i , and ε_t is the random error term. The AR specification provides an estimate of the extent to which the current cardiac state can be explained by its previous observations and therefore served as an interpretable benchmark for comparison with the more comprehensive lagged models.

2.2 Distributed Lag Model

A Distributed Lag Model (DLM) was employed to quantify contemporaneous and delayed relationships between physiological predictors and the cardiac response. The model was expressed as

$$Y_t = \alpha + \sum_{k=0}^q \beta_k X_{t-k} + \varepsilon_t,$$

where X_{t-k} represents the value of a physiological predictor observed k periods before the current cardiac response and β_k denotes the corresponding lag-specific effect. The DLM therefore allowed the study to determine whether the influence of a physiological measurement occurred immediately or persisted across subsequent time points.

2.3 Vector Autoregressive Model

Because cardiac physiological variables may influence one another dynamically, a Vector Autoregressive (VAR) model was used to evaluate multivariate temporal relationships. The VAR(p) model was specified as

$$Y_t = c + \sum_{i=1}^p A_i Y_{t-i} + \varepsilon_t,$$

where Y_t is a vector of cardiac and related physiological variables at time t , c is a vector of intercepts, A_i is the coefficient matrix associated with lag i , and ε_t is a vector of innovations. The VAR framework was used to determine whether previous observations of one physiological variable contributed to subsequent changes in another.

2.4 Regularized Lagged Regression

Including several lagged predictors can introduce multicollinearity and increase the risk of overfitting. To address these problems, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net regularization were incorporated into the analytical framework.

For LASSO, parameters were estimated by minimizing

$$\hat{\beta}_{\text{LASSO}} = \arg \min_{\beta} \left\{ \sum_{t=1}^T \left(Y_t - \beta_0 - \sum_{j=1}^m \beta_j X_{jt} \right)^2 + \lambda \sum_{j=1}^m |\beta_j| \right\},$$

where $\lambda \geq 0$ is the regularization parameter controlling coefficient shrinkage. The L_1 penalty permits some coefficients to be reduced to zero and thereby provides a mechanism for selecting influential lagged predictors.

Elastic Net combines L_1 and L_2 penalties and was specified as

$$\hat{\beta}_{\text{EN}} = \arg \min_{\beta} \left\{ \sum_{t=1}^T \left(Y_t - \beta_0 - \sum_{j=1}^m \beta_j X_{jt} \right)^2 + \lambda \left[\alpha \sum_{j=1}^m |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^m \beta_j^2 \right] \right\},$$

where λ controls the overall degree of regularization and $0 \leq \alpha \leq 1$ determines the relative contribution of the LASSO and ridge penalties. Elastic Net was included because consecutive physiological lags may be highly correlated and can therefore benefit from combined L_1 and L_2 regularization.

2.5 Hybrid Lagged Regression Model

The principal methodological contribution of this study was the development of a Hybrid Lagged Regression Model (HLRM) designed to integrate complementary characteristics of continuous cardiac time-series data within a unified statistical framework. The general HLRM was specified as

$$Y_t = \beta_0 + \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=1}^m \sum_{k=0}^{q_j} \beta_{jk} X_{j,t-k} + \sum_{r=1}^R \gamma_r g_r(\mathbf{X}_t) + \varepsilon_t.$$

Here, Y_t is the cardiac response at time t ; β_0 is the intercept; Y_{t-i} represents autoregressive terms describing persistence in the cardiac response; $X_{j,t-k}$ represents lagged values of physiological predictor j ; β_{jk} is the corresponding distributed-lag coefficient; $g_r(\mathbf{X}_t)$ denotes selected nonlinear transformations or interaction functions; γ_r represents the associated nonlinear parameters; and ε_t denotes the random error term.

To control model complexity and reduce the influence of uninformative lagged predictors, the HLRM was estimated using a penalized objective function of the form

$$\hat{\theta} = \arg \min_{\theta} \left\{ \frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t(\theta))^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\},$$

where θ denotes the full parameter vector, $\hat{Y}_t(\theta)$ is the fitted cardiac response, $\|\beta\|_1$ is the L_1 penalty, and $\|\beta\|_2^2$ is the squared L_2 penalty. Conceptually, the HLRM combines four components: autoregressive persistence, distributed lag effects, regularization, and mild nonlinear effects. The integration of these components was intended to provide greater flexibility than any individual conventional time-series model while retaining explicit parameter interpretation.

2.6 Model Performance Evaluation

Predictive performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), defined respectively as

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2,$$

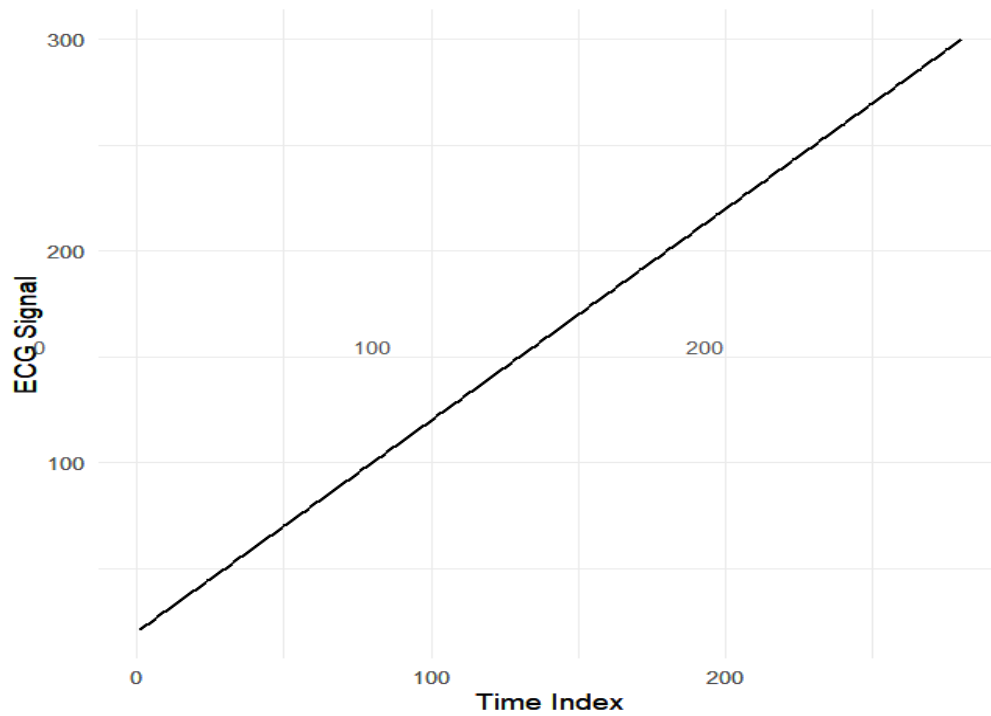
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2},$$

and

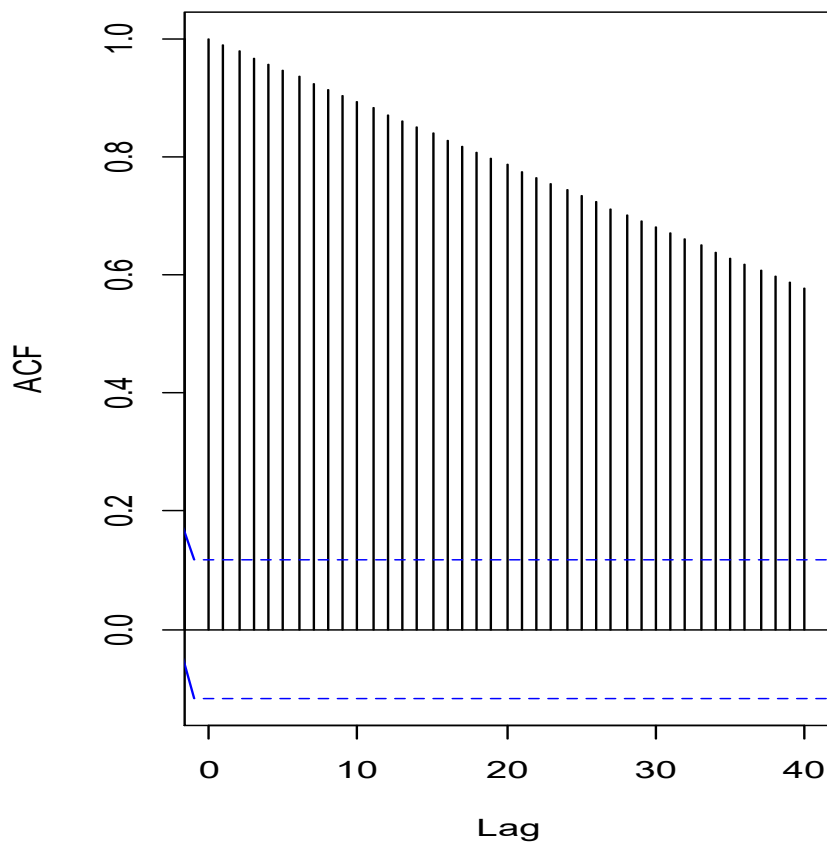
$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^n \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|.$$

Lower values of MSE, RMSE, and MAPE indicate better predictive performance. The HLRM was compared with AR, DLM, VAR, LASSO, and Elastic Net models to determine whether integrating complementary temporal-modeling components improved predictive accuracy.

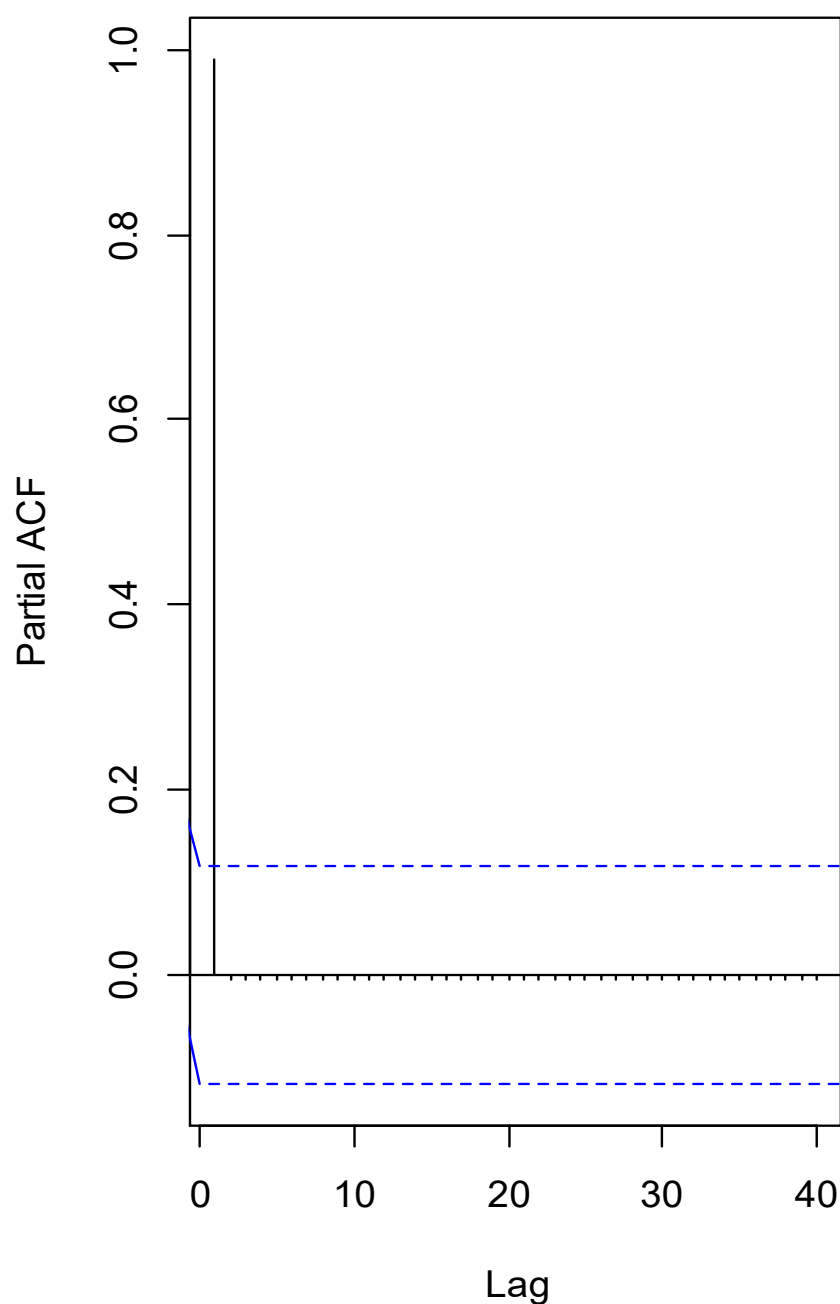
III. Results



Time plot of the ECG signal from Heart Signals



ACF plot of the ECG series from Heart signals



PACF plot of the ECG series from Heart signals

3.1 Hybrid Lagged Regression Model

The estimated HLRM contained significant autoregressive, distributed lag, and nonlinear components. The autoregressive component had the largest coefficient (0.593), followed by the distributed lag component (0.331) and nonlinear component (0.184).

Table 1. Estimated components of the HLRM

Component	Coefficient	Reported significance
Autoregressive	0.593	***
Distributed lag	0.331	***
Nonlinear	0.184	**

The autoregressive component indicates a substantial contribution from previous cardiac observations, while the distributed lag coefficient reflects delayed effects from explanatory physiological variables. The nonlinear term additionally suggests that the cardiac response was not entirely explained by purely linear temporal relationships. The HLRM achieved an R^2 of 0.742 and an adjusted R^2 of 0.718, indicating that approximately 74.2% of the observed variation in the modeled ECG response was explained by the combined model components.

3.2 Predictive Performance Comparison

Predictive accuracy differed consistently across the competing models. The AR model recorded an MSE of 0.046, RMSE of 0.214, and MAPE of 12.8%. Prediction errors decreased progressively for the DLM, VAR, and LASSO models. The proposed HLRM achieved the lowest values for all three error measures, with MSE = 0.024, RMSE = 0.155, and MAPE = 7.9%.

Table 2. Predictive performance of the competing models

Model	MSE	RMSE	MAPE (%)
AR	0.046	0.214	12.8
DLM	0.039	0.197	11.4
VAR	0.034	0.184	10.2
LASSO	0.031	0.176	9.6
HLRM	0.024	0.155	7.9

Relative to the AR benchmark, the HLRM reduced the MSE from 0.046 to 0.024 and the RMSE from 0.214 to 0.155. The consistent reduction across MSE, RMSE, and MAPE indicates improved predictive performance of the hybrid specification compared with the individual models considered.

3.3 Anomaly-Detection Performance

The anomaly-detection performance of the HLRM was assessed using sensitivity, specificity, precision, recall, and the area under the receiver operating characteristic curve (AUC). The model achieved a sensitivity of 0.89 and recall of 0.89, indicating that 89% of anomalous observations were correctly identified. Specificity was 0.84, indicating that 84% of normal observations were correctly classified. Precision was 0.81, while the AUC was 0.92.

Table 3. Anomaly-detection performance of the HLRM

Metric	Value
Sensitivity	0.89
Specificity	0.84
Precision	0.81
Recall	0.89
AUC	0.92

The AUC of 0.92 indicates strong discrimination between anomalous and normal cardiac observations. Taken together with the predictive-error results, the findings show that the HLRM provided the strongest reported performance among the statistical approaches evaluated in the study.

IV. Conclusion

This study developed a Hybrid Lagged Regression Model (HLRM) for modeling continuous cardiac time-series data and supporting the early detection of cardiac anomalies. By integrating autoregressive persistence, distributed lag effects, regularization, and nonlinear components, the proposed framework was designed to capture multiple temporal characteristics of cardiac signals within a unified and interpretable statistical model. The findings showed that the HLRM achieved lower prediction errors than the AR, DLM, VAR, and regularized models considered in the study, while also demonstrating strong reported anomaly-discrimination performance. The study highlights the potential value of combining complementary time-series techniques rather than relying on individual models to characterize complex cardiac dynamics. In particular, the HLRM provides a framework in which temporal persistence, delayed physiological relationships, predictor selection, and nonlinear effects can be examined while maintaining statistical interpretability. However, the findings should be considered preliminary until the model is validated using larger, independent, and more diverse cardiac datasets. Future studies should also investigate multichannel ECG signals, alternative nonlinear structures, and prospective validation. Overall, the HLRM provides a promising methodological foundation for further research on interpretable continuous cardiac anomaly detection.

References

- [1]. Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: A review. *Medical & Biological Engineering & Computing*, 44(12), 1031–1051. <https://doi.org/10.1007/s11517-006-0119-0> (PubMed)
- [2]. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
- [3]. Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.
- [4]. Clifford, G. D., Azuaje, F., & McSharry, P. E. (Eds.). (2006). *Advanced methods and tools for ECG data analysis*. Artech House.
- [5]. Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.
- [6]. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65–69. <https://doi.org/10.1038/s41591-018-0268-3> (Nature)
- [7]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- [8]. Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, Article 1. <https://doi.org/10.1038/s41597-022-01899-x> (PubMed)
- [9]. Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.
- [10]. Oyeleye, M., Chen, T., Titarenko, S., & Antoniou, G. (2022). A predictive analysis of heart rates using machine learning techniques. *International Journal of Environmental Research and Public Health*, 19(4), Article 2417. <https://doi.org/10.3390/ijerph19042417> (PubMed Central (PMC))
- [11]. Popovska, E., & Georgieva-Tsaneva, G. (2024). Analysis and prediction of time-series data heart rate variability using ARIMA model. *Science Series—Innovative STEM Education*, 6, 225–235.
- [12]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x> (Nature)
- [13]. Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples* (4th ed.). Springer. This is also the edition recorded in your dissertation.
- [14]. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x> (Wiley Online Library)
- [15]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x> (Wiley Online Library)