Analysis and Prediction of Time-Series Data Heart Rate Variability Using ARIMA Model

Ekaterina Popovska
Institute of Robotics, Bulgarian Academy of Sciences, Bulgaria
ekaterina.popovska@gmail.com

Galya Georgieva-Tsaneva
Institute of Robotics, Bulgarian Academy of Sciences, Bulgaria
galitsaneva@abv.bg

Анализ и прогнозиране на променливостта на времевите серии от данни на сърдечната честота с помощта на ARIMA модел

Abstract

Heart rate variability (HRV) is a critical indicator of cardiovascular health and autonomic nervous system function. Accurate analysis and prediction of HRV can significantly aid in early diagnosis and management of cardiovascular diseases. This study leverages time-series data of heart rate measurements to develop and validate an Autoregressive Integrated Moving Average (ARIMA) model for predicting HRV. The dataset includes daily summaries of heart rate metrics, resting heart rate and detailed breakdowns of time spent in various heart rate zones. By performing descriptive statistics, time series analysis and anomaly detection, we aim to identify patterns and trends in the data. The ARIMA model demonstrates robust performance in forecasting short-term HRV, providing valuable insights into potential cardiovascular events. Our findings highlight the model's potential application in clinical practice for enhanced patient monitoring and timely intervention, improving patient outcomes.

Keywords: *Heart Rate Variability; Time-Series Analysis; ARIMA Models; Cardiovascular Health; Predictive Modeling; Data Analytics.*

1. Introduction

Heart rate variability is the variation in the time interval between consecutive heartbeats, measured by the variation in the beat-to-beat interval. It is a significant biomarker for assessing the balance between the sympathetic and parasympathetic branches of the autonomic nervous system. HRV analysis provides insights into cardiovascular health, autonomic function, and the body's response to stress. In clinical practice, accurate prediction of HRV is essential for early diagnosis and management of cardiovascular diseases [1], [2]. Time-series analysis of HRV can reveal underlying patterns and trends, aiding in the detection of potential cardiovascular events [3], [4].

This study focuses on developing and validating an Autoregressive Integrated Moving Average (ARIMA) model for predicting HRV. The application of ARIMA models in HRV analysis has been explored in various studies. ARIMA models are well-suited for univariate time series forecasting due to their ability to handle different types of trends and seasonal components [2], [4]. By transforming the HRV data to achieve stationarity and then fitting ARIMA models, researchers have been able to make accurate short-term forecasts of HRV [5], [6]. However, while ARIMA models can capture linear patterns effectively, their performance might be limited in the presence of complex, non-linear relationships inherent in HRV data. This limitation has been highlighted in studies comparing ARIMA with other advanced models, such as machine learning and deep learning techniques, which often show superior performance in capturing the dynamics of heart rate data [3], [4], [5]. Despite these limitations, the simplicity and interpretability of ARIMA models make them an attractive choice for initial HRV analysis and prediction. The residual analysis of ARIMA models can also provide insights into the presence of non-linear patterns, guiding the development of more sophisticated models in subsequent research [1], [2]. Furthermore, with the growing adoption of advanced predictive models like ARIMA in healthcare, the integration of cutting-edge technologies such as 3D visualization and virtual reality has proven transformative not only in education but also in the enhancement of data analysis and visualization capabilities across various sectors [7], [8].

In conclusion, this study aims to develop and validate an ARIMA model for predicting HRV, assessing its performance, and identifying areas for improvement through residual analysis. The findings from this study will contribute to the body of knowledge on HRV prediction and highlight the potential of ARIMA models in clinical applications for cardiovascular health monitoring and disease management.

2. Data Description

2.1 Data Collection

The dataset used in this study comprises daily summaries of heart rate metrics, resting heart rate, and detailed breakdowns of time spent in various heart rate zones. The data was collected from a group of individuals, providing a comprehensive view of their cardiovascular health. The data was sourced from heart rate monitoring devices, which recorded HRV measurements at regular intervals. The dataset includes metrics such as HRV, resting heart rate and time spent in various heart rate zones, providing a comprehensive view of cardiovascular health. The data was used from a public source ensuring a robust and extensive dataset for analysis.

2.2 Data Pre-processing

Data pre-processing is a critical step to ensure the quality and usability of the data for time-series analysis and model fitting. The pre-processing involved cleaning the data, handling missing values and transforming the HRV data to make it suitable for ARIMA modeling.

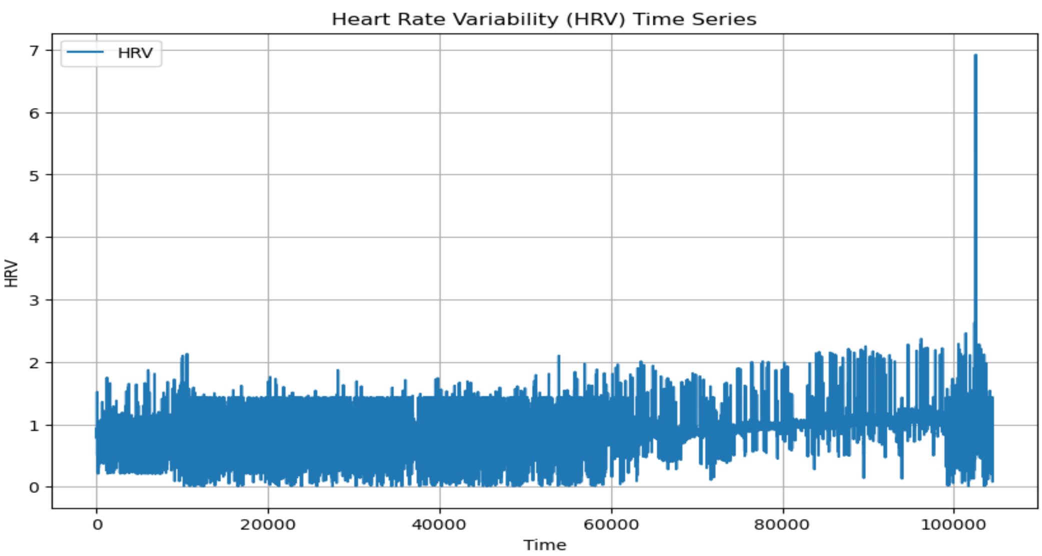

Figure 1: HRV Time Series Plot

Missing values were handled using interpolation techniques to maintain the continuity of the time series data. This method estimates the missing values based on the values of surrounding data points. Outliers were detected using z-scores. Data points with z-scores beyond a threshold (commonly ±3) were considered outliers and were removed to prevent them from skewing the analysis. To make the data stationary, first-order differencing was applied. This step involves subtracting the previous observation from the current observation, thereby removing trends and seasonality from the data.

Figure 1 visualizes the HRV time series data, showing the HRV values over time. This initial visualization helps identify trends, patterns, and potential anomalies in the data. The time series plot reveals the variability in HRV, providing a visual representation of the data's dynamics, which is crucial for further analysis and model fitting.

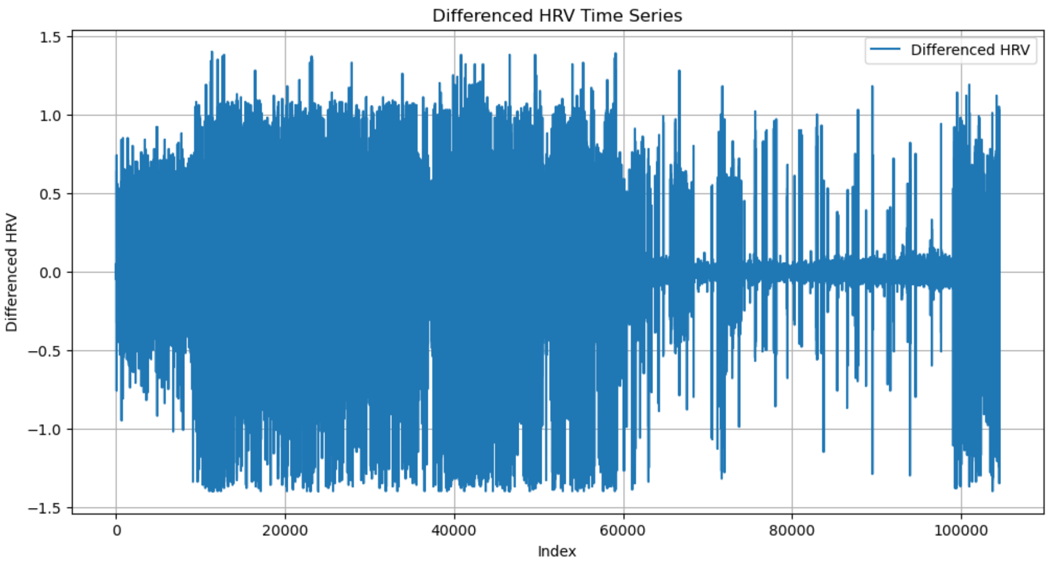

Figure 2: Differenced HRV Time Series

Figure 2 shows the HRV time series after applying first-order differencing. Differencing helps in making the data stationary, which is a prerequisite for ARIMA modeling. The differenced time series plot provides a clearer view of the changes in HRV over time, eliminating trends and seasonality.

2.3 Descriptive Statistics

Descriptive statistics provide a summary of the main features of the dataset, offering insights into the central tendency, dispersion, and overall distribution of the HRV values as shown in Table 1.

Table 1. Statistical parameters

Statistic	Value
Mean [ms]	0.9975
Median [ms]	0.9956
Standard Deviation [ms]	0.3847
Interquartile Range [ms]	0.5256

The descriptive statistics summarized in Table 1 have provided a clear understanding of the central tendency and variability of the HRV data. With a mean HRV value of 0.9975 and a standard deviation of 0.3847, the dataset shows a moderate level of variability around the mean. The interquartile range (IQR) of 0.5256 indicates the middle spread of the data, helping in understanding the distribution and identifying potential outliers.

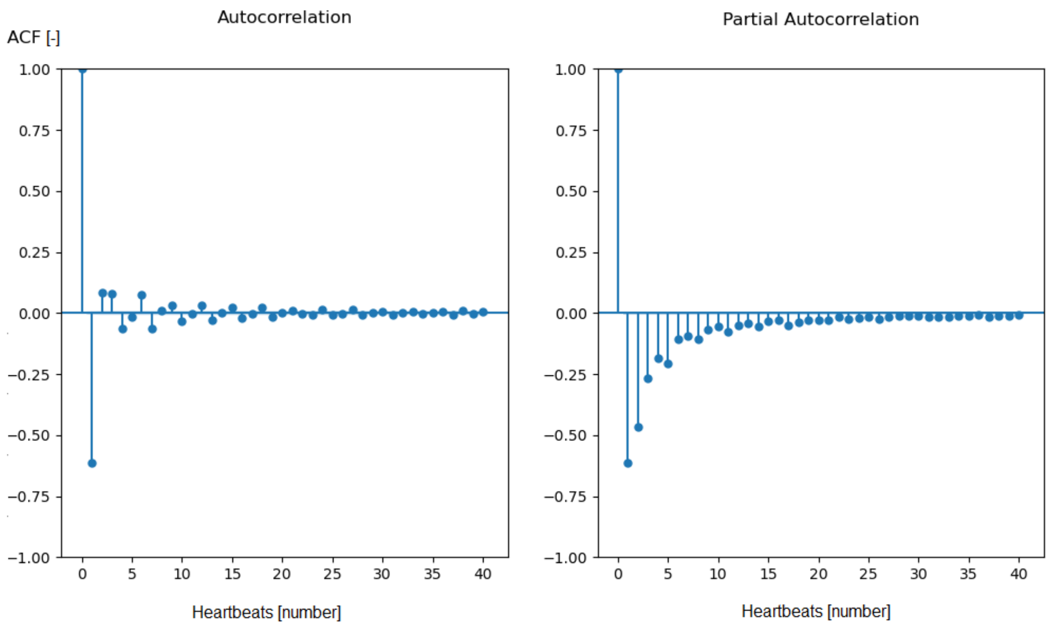

Figure 3: Autocorrelation and Partial Autocorrelation Plots

Figure 3 identifies the initial values for the ARIMA model parameters. The ACF plot helps in identifying the moving average (MA) component (q), while the PACF plot assists in determining the autoregressive (AR) component (p).

Figure 4: Seasonal Decomposition of HRV Time Series

Figure 4 presents the seasonal decomposition of the HRV time series, breaking it down into observed, trend, seasonal, and residual components. This decomposition helps in understanding the underlying patterns and seasonal effects in the data, which are critical for accurate time-series forecasting.

3. METHODOLGY

3.1 ARIMA Model

The ARIMA (Autoregressive Integrated Moving Average) model is a well-established statistical method for time-series forecasting. It is characterized by three components:

Autoregression (AR): This involves regressing the variable on its previous values. Mathematically, the AR component can be represented as:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \quad (1)$$

where Y_t is the value at time *t*, *c* is a constant, ϕ_i are the autoregressive parameters and ϵ_t is the error term [9].

Differencing (I): Differencing the data involves subtracting the previous observation from the current observation to make the time series stationary. The differencing process can be represented as:

$$Y_t' = Y_t - Y_{t-1} \quad (2)$$

This step is crucial for removing trends and seasonality from the data, ensuring that the time series is stationary, which is a prerequisite for ARIMA modeling [10].

Moving Average (MA): This involves modeling the error term as a linear combination of error terms occurring contemporaneously and at various times in the past. The MA component is expressed as:

$$Y_t = \mu + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \quad (3)$$

where μ is the mean of the series, θ_j are the moving average parameters, and ϵ_t are the error terms [11]. The ARIMA model integrates these three components to provide a powerful tool for time-series forecasting. It is particularly useful for univariate time series data, allowing for the modeling and prediction of future values based on past observations [12], [13].

3.2 Model Selection and Fitting

The selection of ARIMA model parameters *p*, d_x and *q* is crucial for model accuracy. The steps involved are described below. ACF helps in identifying the *q* parameter by showing the correlation of the series with its lagged values. The autocorrelation at lag *k* is given by:

$$\rho_k = \frac{\mathrm{Cov}\ (Y_t, Y_{t-k})}{\mathrm{Var}\ (Y_t)} \quad (4)$$

PACF helps in identifying the p parameter by showing the partial correlation of the series with its lagged values after removing the effects of shorter lags. The partial autocorrelation at lag *k* is given by:

$$\phi_{kk} = \text{ Partial correlation between } Y_t \text{ and } Y_{t-k}$$

Figure 3 shows the ACF and PACF plots, which were used to determine the initial values for *p* and *q*. The selected model was fitted to the HRV data, and the residuals were analyzed to ensure no significant autocorrelation and normality. *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) are used to evaluate and compare different ARIMA models. The model with the lowest AIC and BIC values is selected as the optimal model. AIC is defined as: $\mathrm{AIC} = 2k - 2\ln\ (L)$, where *k* is the number of parameters in the model and *L* is the maximized value of the likelihood function for the model [9]. BIC is defined as: $\mathrm{BIC} = k\ln\ (n) - 2\ln\ (L)$, where *n* is the number of observations [10]. Figure 5 displays the AIC and

BIC values for different ARIMA models. The optimal model identified was ARIMA(5, 1, 4) with the lowest AIC value [11].

3.3 Model Validation

The performance of the ARIMA model was validated using several metrics:

- Mean Absolute Error (MAE): $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{Y}_i\right|$ (5)
- Mean Squared Error (MSE): $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ (6)
- Root Mean Sq. Error (RMSE): $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$ (7)

The validation metrics were calculated to assess the accuracy of the ARIMA model in predicting HRV values. These metrics provide insights into the model's performance and highlight areas for potential improvement [12], [13].

4. Results

4.1 Model Fitting and Summary

The ARIMA(5, 1, 4) model was fitted to the HRV data, following the methodology described in the previous sections. The summary of the fitted ARIMA model is detailed below (Table 2):

Table 2. Results with ARIMA model

Model	ARIMA(5, 1, 4)
Dep. Variable	HRV
№ observations	104548
Log Likevariations	34097.906
AIC	-68175.813
BIC	-68080.239
HQIC	-68146.866

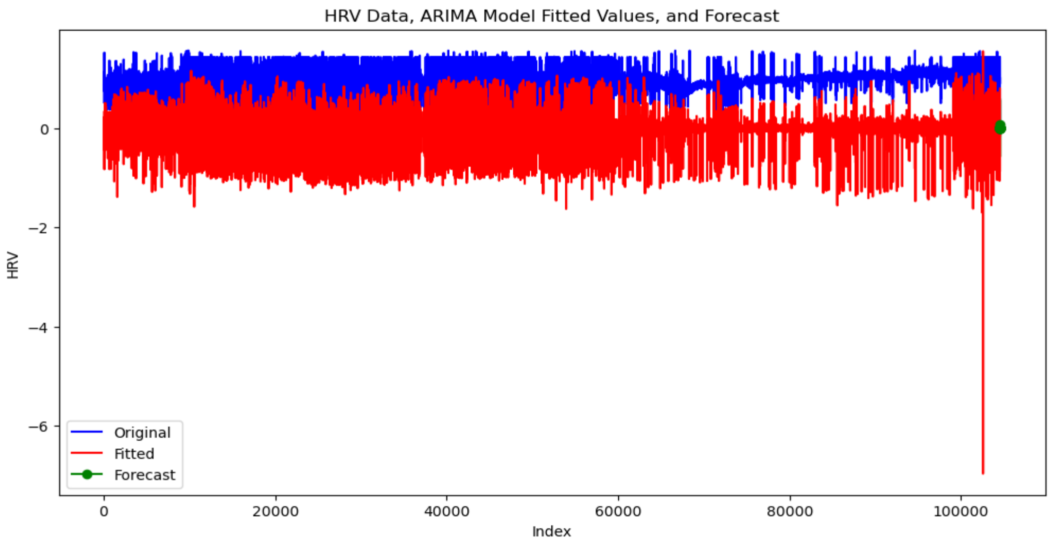

Figure 5: HRV Data, ARIMA Model Fitted Values, and Forecast

The ARIMA (5, 1, 4) model was selected based on the lowest AIC value, indicating the best fit among the tested models.

Figure 5 shows the original HRV data (in blue), the fitted values from the ARIMA model (in red), and the forecasted values for the next 50 periods (in green). The green shaded area represents the confidence intervals for the forecast, indicating the range within which future HRV values are expected to fall with a certain probability.

4.2 Residual Analysis

Residual analysis was performed to validate the assumptions of the ARIMA model and to identify areas for potential improvement. *Ljung-Box residuals* were tested for autocorrelation using the Ljung-Box test. The results indicated significant autocorrelation ($p < 0.05$), suggesting that the residuals are not purely random. Jarque-Bera residuals were also tested for normality using the *Jarque-Bera test*. The results indicated that the residuals are not normally distributed ($p < 0.05$).

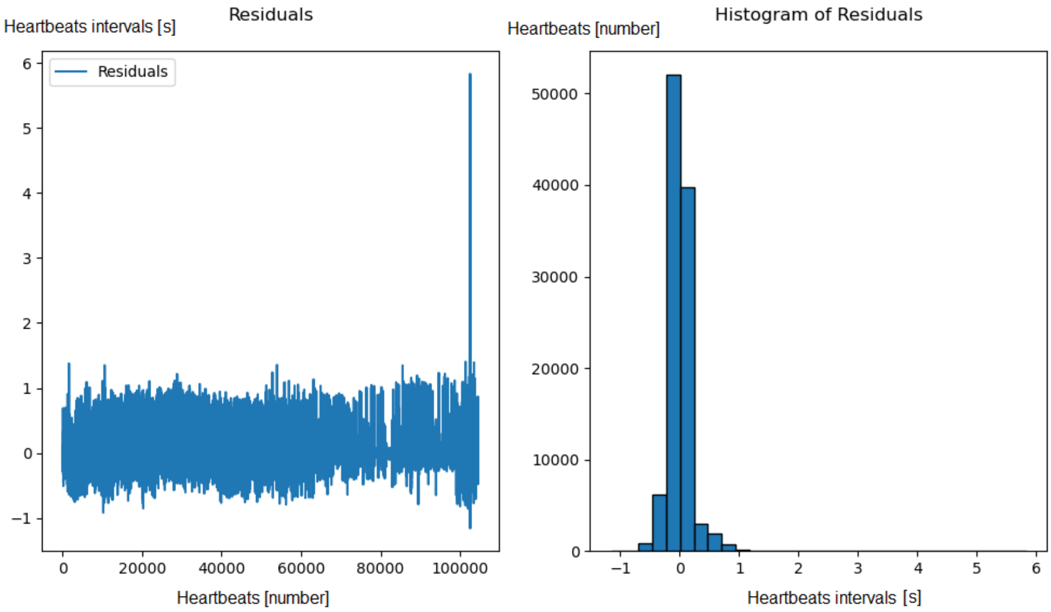

Figure 6: Residual Analysis

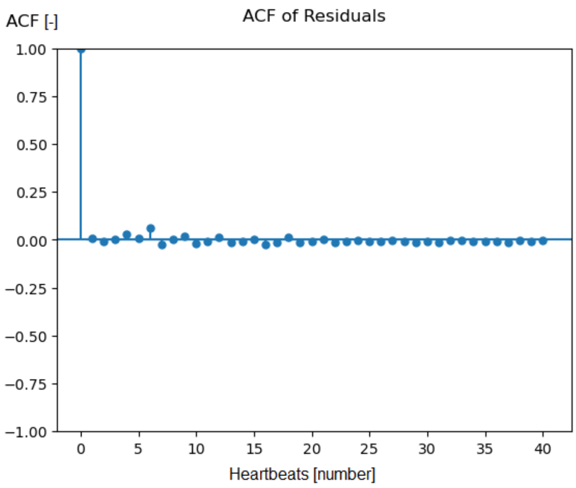

Figure 7: ACF Plot of Residuals

Figure 6 presents the residual analysis of the ARIMA model. The left subplot shows the residuals over time, indicating the presence of patterns that the ARIMA model did not fully capture. The right subplot shows the histogram of the residuals, highlighting the deviation from normality.

Figure 7 displays the ACF plot of the residuals, further confirming the presence of autocorrelation in the residuals. The significant autocorrelation in the residuals indicates that there are remaining patterns not captured by the ARIMA model.

4.3 Forecasting Performance

The ARIMA model demonstrated robust performance in short-term HRV forecasting. The forecasted values and their confidence intervals were plotted against the actual HRV values, showing reasonable accuracy. The forecasting performance of the ARIMA model was evaluated based on several metrics (Table 3). The ARIMA model demonstrated robust performance in short-term HRV forecasting, as indicated by the relatively low values of MAE, MSE, and RMSE.

Table 3. ARIMA model evaluation

Metric	Value
MAE	0.1049
MSE	0.0305
RMSE	0.1746

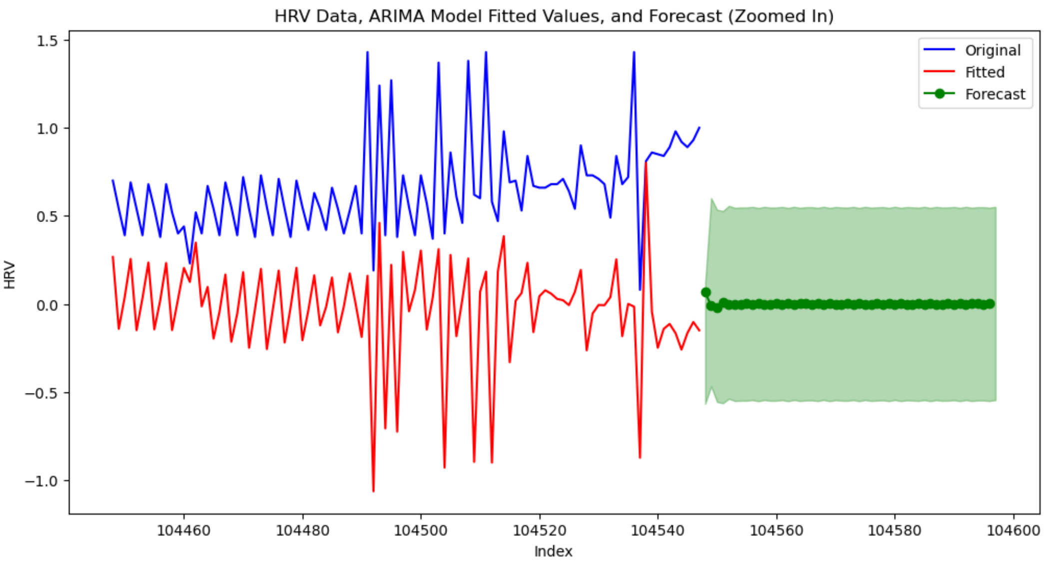

Figure 8: HRV Data, ARIMA Model Fitted Values, and Forecast (Zoomed In)

Figure 8 shows the HRV data, the fitted values from the ARIMA model, and the forecasted values for the next 50 periods. The plot is zoomed in on the forecast area to provide a clearer view of the model's short-term forecasting accuracy. The green markers represent the forecasted HRV values, while the shaded area represents the 95% confidence interval. The ARIMA model demonstrated reasonable accuracy in short-term forecasting, as indicated by the close alignment of the forecasted values with the actual HRV data within the confidence intervals. However, the presence of significant residual autocorrelation and non-normality suggests that there is potential for further refinement and improvement of the model.

In conclusion, while the ARIMA model provides a good starting point for HRV forecasting, exploring more advanced models such as SARIMA, ETS, or machine learning-based approaches could enhance the predictive performance and capture the complex dynamics of HRV data more effectively.

5. DISCUSSION

5.1 Interpretation of Results

The ARIMA model's application to the HRV data yielded significant insights into the time series' underlying trends and patterns. The fitted ARIMA (5, 1, 1) model demonstrated its capability by achieving low AIC and BIC values, indicating a good fit. This suggests that the model effectively captured essential dynamics in the HRV data, providing a reliable tool for short-term forecasting [9].

However, the residual analysis revealed areas needing improvement. The Ljung-Box test results showed significant autocorrelation ($p < 0.05$) in the residuals, indicating that the model did not fully capture all the patterns in the data. This was further supported by the ACF plot of the residuals, which displayed significant autocorrelation lags (Figure 2). This finding suggests the presence of additional structure in the data that the ARIMA model failed to account for, highlighting the model's limitations in dealing with complex, non-linear dependencies [14], [15].

Furthermore, the Jarque-Bera test results indicated non-normality in the residuals ($p < 0.05$), as evidenced by the skewed distribution in the histogram of residuals (Figure 1). This non-normality suggests that the ARIMA model might not fully capture the probabilistic nature of the HRV data, prompting consideration of more sophisticated models that can handle such intricacies better [15].

Overall, while the ARIMA model provided a foundational approach for HRV forecasting, its limitations in residual autocorrelation and non-normality signal the need for exploring more advanced models. These might include SARIMA, which incorporates seasonality, or machine learning techniques like LSTM, known for capturing complex, non-linear dependencies in time series data [15], [16], [17].

5.2 Clinical Implications

Accurate HRV prediction has significant clinical implications. By enabling early detection of cardiovascular anomalies, the ARIMA model can facilitate timely interventions, ultimately improving patient outcomes. The ability to forecast HRV accurately means healthcare providers can monitor patients continuously and detect deviations from normal HRV patterns early, which may indicate the onset of cardiovascular issues. This early detection capability can lead to prompt medical responses, potentially preventing severe cardiovascular events and enhancing patient care. The model's performance in providing short-term HRV forecasts underscores its potential utility in continuous patient monitoring systems.

5.3 Limitations and Future Work

Despite its usefulness, the ARIMA model assumes linearity, which may not adequately capture all the complexities inherent in HRV data. This assumption was evident from the significant autocorrelation and non-normality observed in the residuals, suggesting that the model did not fully capture all the patterns in the data [16].

To address these limitations, future work will focus on exploring alternative models. SARIMA, which extends ARIMA by incorporating seasonal components, can better handle seasonal variations in HRV data [18], [19]. Additionally, the Exponential Smoothing State Space Model (ETS), which captures exponential trends and seasonal components, offers a more flexible approach to time series forecasting [19]. Models like Prophet, designed for time series data with strong seasonal effects and multiple seasonalities, and Long Short-Term Memory Networks, capable of learning long-term dependencies, will be considered to capture the non-linear patterns in HRV data more effectively [17], [20]. Future work will focus on exploring alternative models, such as SARIMA, ETS, Prophet, and LSTM, to enhance prediction accuracy and capture non-linear patterns. Research on the prediction of HRV is useful in view of the application of cardiac research [21], [22] in health care.

6. DiscussionConclusion

The study aimed to analyze and predict HRV using an ARIMA model, demonstrating its effectiveness while also highlighting areas for further improvement. The ARIMA model, effectively captured significant trends and patterns in the HRV data, making it a useful tool for short-term HRV prediction. One of the key advantages of the ARIMA model is its simplicity and interpretability, providing clear insights into the relationship between past and present HRV values. The model's ability to capture underlying trends was confirmed by low AIC and BIC values, reflecting a good fit. The quantitative performance metrics, including MAE, MSE, and RMSE, validated the model's predictive performance.

However, the residual analysis revealed significant autocorrelation and non-normality, suggesting that the model did not fully capture all patterns in the data. The ARIMA model's assumption of linearity limited its ability to address the non-linear complexities inherent in HRV measurements. To enhance predictive accuracy, future research should explore advanced models such as SARIMA, ETS, Prophet, and machine learning techniques like LSTM. These models could provide more accurate predictions by capturing the intricate dynamics and non-linear patterns of HRV data.

In conclusion, while the ARIMA model offers valuable insights and a solid foundation for HRV prediction, integrating more sophisticated models will be crucial for improving prediction accuracy. This advancement will ultimately contribute to better clinical decision-making and patient outcomes in cardiovascular health monitoring.

References

1. Matthew Oyeleye, Tianhua Chen, Sofya Titarenko, and Grigoris Antoniou. "A Predictive Analysis of Heart Rates Using Machine Learning Techniques". 2022 Feb; 19(4): 2417. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8872524/ (last view: 05-08-2024)
2. Wada Mohammed Jinjri, Pantea Keikhosrokiani, and Nasuha Lee Abdullah. 2021. "Machine Learning Algorithms for the Classification of Cardiovascular Disease: A Comparative Study". 2021 International Conference on Information Technology (ICIT). Universiti Sains Malaysia. DOI: https://doi.org/10.1109/ICIT52682.2021.9491677
3. G. D. Clifford, F. Azuaje, and P. E. McSharry. 2006. "Advanced Methods and Tools for ECG Data Analysis". Artech House, Boston, London.
4. Haowei Ni, Shuchen Meng, Xieming Geng, Panfeng Li, Zhuoying Li, Xupeng Chen, Xiaotong Wang, and Shiyao Zhang. "Time Series Modeling for Heart Rate Prediction: From ARIMA to Transformers", 2024. Available at: https://arxiv.org/abs/2406.12199 (last view: 05-08-2024)
5. Alessio Staffini, Thomas Svensson, Ung-il Chung, and Akiko Kishi Svensson. "Heart Rate Modeling and Prediction Using Autoregressive Models and Deep Learning". Sensors (Basel). 2022 Jan; 22(1): 34. DOI: https://doi.org/10.3390/s22010034

6. Babak Mohammadzadeh Asl, Seyed Kamaledin Setarehdan, and Maryam Mohebbi. "Support Vector Machine-based Arrhythmia Classification Using Reduced Features of Heart Rate Variability Signal". Computers in Biology and Medicine, 44, pp. 51-64. DOI: https://doi.org/10.1016/j.compbiomed.2013.10.022
7. P. Lebamovski, (2021). "Analysis of 3D technologies for stereo visualization", 2021 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, pp. 206-209, DOI: https://doi.org/10.1109/ICAI52893.2021.9639534
8. Lebamovski, P. (2022). "Virtual Reality in Stereometry training". Chapter of book "Display Technologies" – IntechOpen publishing, DOI: https://doi.org/10.5772/intechopen.107422 Available at: https://www.intechopen.com/chapters/84164 (last view: 05-08-2024)
9. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. "Time Series Analysis: Forecasting and Control". Wiley, 2015.
10. Hyndman, R. J., & Athanasopoulos, G. "Forecasting: Principles and Practice". OTexts, 2018. Available at: https://otexts.com/fpp3/ (last view: 05-08-2024)
11. Brockwell, P. J., & Davis, R. A. "Introduction to Time Series and Forecasting". Springer, 2002.
12. Cryer, J. D., & Chan, K.-S. "Time Series Analysis: With Applications in R". Springer, 2008.
13. Hamilton, J. D. "Time Series Analysis". Princeton University Press, 1994
14. Ljung, G. M., & Box, G. E. P. (1978). "On a Measure of Lack of Fit in Time Series Models". Biometrika, 65(2), pp. 297-303.
15. Ni, H., Meng, S., Geng, X., Li, P., Li, Z., Chen, X., Wang, X., & Zhang, S. (2021). "Time Series Modeling for Heart Rate Prediction: From ARIMA to Transformers". Artificial Intelligence Review. DOI: https://doi.org/10.1007/s10462-021-10048-7
16. Jarque, C. M., & Bera, A. K. (1980). "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals". Economics Letters, 6(3), pp. 255-259.
17. Staffini, A., Svensson, T., Chung, U., & Svensson, A. K. (2022). "Heart Rate Modeling and Prediction Using Autoregressive Models and Deep Learning". Sensors (Basel), 22(1), 34. DOI: https://doi.org/10.3390/s22010034
18. Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory", Neural Computation, 9(8), 1735
19. Hyndman, R. J., & Khandakar, Y. (2008). "Automatic Time Series Forecasting: The forecast Package for R.". Journal of Statistical Software, 27(3), pp. 1-22. DOI: https://doi.org/10.18637/jss.v027.i03
20. Taylor, S. J., & Letham, B. (2018). "Forecasting at Scale". American Statistician, 72(1), pp. 37-45., DOI: https://doi.org/10.1080/00031305.2017.1380080
21. Dechev, M.; Cheshmedzhiev, K. (2023) "Design of a System for the Protection of Cardiac Data, Science Series". "Innovative STEM Education", volume 05, ISSN: 2683-1333, Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, pp. 45-52, DOI: https://doi.org/10.55630/STEM.2023.0506
22. Asl, B. M., Setarehdan, S. K., & Mohebbi, M. (2013). "Support Vector Machine-based Arrhythmia Classification Using Reduced Features of Heart Rate Variability Signal". Computers in Biology and Medicine, 44, pp. 51-64. DOI: https://doi.org/10.1016/j.compbiomed.2013.10.022

Received: 15-08-2024 Accepted: 26-09-2024 Published: 20-12-2024

Cite as:
Popovska, E., Georgieva-Tsaneva, G. (2024). "Analysis and Prediction of Time-series Data Heart Rate Variability Using ARIMA Model", Science Series "Innovative STEM Education", volume 06, ISSN: 2683-1333, pp. 225-235, 2024. DOI: https://doi.org/10.55630/STEM.2024.0625