Journal of Healthcare Informatics Research (2025) 9:154–173
https://doi.org/10.1007/s41666-025-00191-y

RESEARCH ARTICLE

Real-Time Forecasting from Wearable-Monitored Heart Rate Data Through Autoregressive Models

Giulio De Sabbata[1] · **Giovanni Simonini**[1]

Received: 13 June 2024 / Revised: 31 January 2025 / Accepted: 10 February 2025 /
Published online: 7 March 2025

Abstract

Heart rate (HR) analysis is of paramount importance in healthcare, particularly for monitoring cardiovascular health, a global concern. The advent of wearable sensors has enabled continuous HR monitoring, with researchers attempting to develop early detection systems by forecasting HR in a univariate fashion. This study analyzes real-world HR time series gathered during participants daily routines to critically assess the predictive power of past HR data in short-term, univariate forecasting. The literature emphasizes a minute-by-minute, univariate forecasting approach, where state-of-the-art predictive models predominantly employ autoregressive integrated moving average (ARIMA). Yet, its superiority has been proved without studying its optimized hyper-parameters, which could not only improve forecast accuracy but also provide valuable insights. By leveraging the interpretability of ARIMA, we tune its hyper-parameters within a minute-by-minute forecasting structure to address the central research question: how does historical HR data contribute to generate accurate short-term HR forecasts? Our analysis finds that the random walk model, a special case of ARIMA, consistently performs comparably to, or even better than, more complex ARIMA specifications. This indicates that HR values alone offer limited predictive power for short-term forecasting, casting doubt on the value of further refinement in univariate models for alarm system development. These findings highlight the limitations of univariate HR forecasting in real-time health monitoring. Rather than increasing model complexity, future research might benefit from exploring alternative approaches to improve early warning system capabilities in real-world settings.

Keywords Heart rate · Forecasting · Wearable sensors · Real-time · Random walk

✉ Giulio De Sabbata
giulio.desabbata@unimore.it

Giovanni Simonini
giovanni.simonini@unimore.it

[1] University of Modena and Reggio Emilia, Modena, Italy

1 Introduction

The digitization of healthcare has transformed the sector, offering researchers a wealth of new disruptive opportunities. Among these, the proliferation of wearable sensors has generated a huge amount of electronic health data and, thus, fostered further research on one of the major vital signs, the heart rate (HR). HR analysis is crucial in healthcare, especially for the study of cardiovascular diseases, a major global health concern that demands particular attention. Indeed, these diseases continue to be a leading cause of mortality worldwide [1]. Beyond the sheer volume of data, wearable sensors have facilitated real-time monitoring of various physiological parameters in patients throughout their daily routines. In practical healthcare settings, monitoring and alarm systems must operate in real-time, ensuring timely interventions when necessary [2]. The wealth of new studies collecting HR time series has encouraged researchers to elicit insights from this new source of information to develop early detection system [3–10]. Among these, the prevailing approach in recent HR works has leaned towards univariate forecasting of HR values, employing a short-term forecasting strategy with a horizon of 1 min.

Under the assumption that accurate prediction is sufficient for developing effective alarm systems, numerous machine learning and complex deep learning models have been tested to generate real-time forecasts. However, this approach represents a considerable leap, as the efficacy of these systems remains unproven. A primary obstacle remains the scarcity of labeled data in longitudinal studies. This limitation stems from the nature of the available datasets. Studies typically collect data either from healthy individuals or from patients already diagnosed with cardiovascular conditions. Consequently, capturing the precise onset of cardiac events remains elusive. Moreover, the field progress appears to be driven more by comparative analyses of algorithm superiority rather than in view of early detection systems. This focus on algorithmic competition may yield limited benefits in terms of clinical applications.

Given these challenges, we propose that a critical starting point for advancing the field is to assess the performance of univariate forecasting models. This leads us to a central research question: to what extent do predictive models leverage past information in univariate HR forecasting? To address this question, we aim at identifying the most performing specification of autoregressive integrated moving average (ARIMA), which is a highly interpretable model that has shown promising performance compared to other predictive models [5, 6]. By tuning ARIMA hyper-parameters, a rigorous evaluation of model specifications allows us to examine the predictive power of past HR data and assess the potential of univariate models in capturing meaningful patterns. Notably, finance literature emphasizes the role of the random walk model, represented by ARIMA (0,1,0), using it as a benchmark [11]. This derives from the nature of the random walk model, where forecasts are generated by taking a random step away from the previous value, meaning that such a model does not retrieve information from lagged and error term. To assess how past information is leveraged to predict future values, performance of other specifications are compared to that of a random walk model. If the random walk model provides a competitive, if not superior, forecasting baseline, it could be concluded that past information is minimally leveraged. Our findings may help clarify the limitations and potential of univariate HR forecasting,

guiding future research towards more effective approaches in cardiovascular health monitoring. Once ARIMA specification is tuned, analysis of residuals is performed to deepen our understanding of model behavior.

Identifying the best specification of hyper-parameters drastically reduces the complexity time of algorithm deployment. Time constraints should be considered when deploying real-time solutions [12, 13]. Further tests are necessary to assess the possibility of deploying the solution in production. While these tests are not covered in this paper, it should be noted that our methodology designed a solution tailored to real-time scenarios. The main contributions of the paper can be summarized as follows:

- `Established a predominant forecasting approach.`
  Outlined key passages for designing a forecasting structure in the HR domain, enhancing comparability and comprehension among researchers. Notions of granularity, adaptive learning, and forecasting strategy are crucial for the definition of the forecasting structure;
- `ARIMA hyper-parameters testing.`
  This study is the first to conduct a comprehensive testing of different specifications of ARIMA for HR forecasting. Hyper-parameter tuning provides valuable insights into the forecast accuracy associated to historical HR values;
- `Assessment of predictive performance in univariate forecasting.`
  We evaluate how historical information is leveraged to generate short-term forecasts in HR univariate setting;
- `Implementation of residual diagnostics in the random walk model.`
  Performance of the random walk model is evaluated through residual diagnostics, which represents a means of studying model behavior.

The remainder of the paper is organized as follows: Section 2 deals with the methodology followed to design our minute-by-minute forecasting structure. In particular, a description of the datasets is provided underscoring its adequacy for our purpose. Subsequently, data pre-processing phase is debated focusing on the organization of the dataset and related cleaning tasks. The definition of adaptive learning strategies is further inspected. Lastly, ARIMA algorithm, validation process, and evaluation metrics are analyzed in this section. Section 3 presents results of the tests with a description of the experimental protocol and of the relative residual diagnostics. Section 4 summarizes the findings and relative interpretations. Section 5 discusses the literature, especially in the context of HR and time series forecasting. Section 6 reflects on the possible impact on further research in the field.

2 Methods

2.1 Real-World Data: Participants Engaged in Everyday Routines

For this study, we employ two longitudinal datasets, each providing continuous 24-h data gathered through wearable devices. Using wearable sensors to collect data

allows to capture a wide range of HR patterns, thereby mirroring real-world variability. Both datasets consist of healthy individuals engaged in their routine daily activities, confirming comprehensive representation of HR dynamics in natural settings.

The first dataset, referred to as MMASH (Multi-Modal Ambulatory Stress and Heart Rate), is titled *A Public Dataset of 24-h Multi-Levels Psycho-Physiological Responses in Young Healthy Adults* [14]. It contains 24 h of continuous psycho-physiological data, e.g., inter-beat interval data, HR data, wrist accelerometry data, sleep quality index, and physical activity. The 22 participants are young healthy males that are monitored during their normal routines.

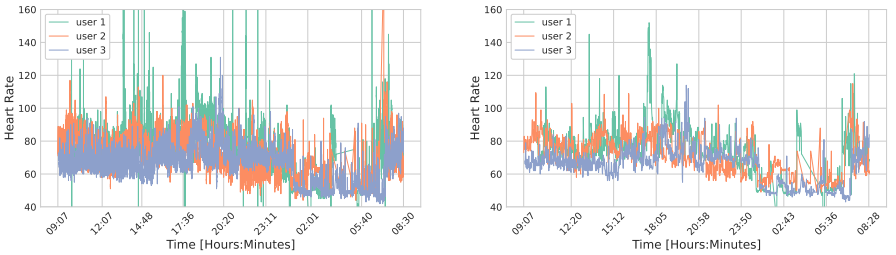
The second dataset, named *RR interval time series from healthy subjects* (RRITS), encompasses 24-h Holter monitoring data from 147 individuals [15]. The dataset is nearly gender-balanced and includes participants aged 0 to 55, though skewed towards those under 1 year old. While subject-specific factors such as age and gender may affect outcomes in univariate analyses, the inclusion of this heterogeneous sample leads to findings that are more generalizable and less influenced by individual demographic factors.

2.2 Pre-Processing in Heart Rate Time Series

In both datasets, a wearable device is employed to collect data that, thus, may be affected by motion artifacts. Unlike conventional instruments commonly used in clinical settings for HR monitoring, wearable sensors are more prone to generating inaccuracies and errors in the measurements. In MMASH and RRITS, HR data is collected using a PolarH7 sensor and a Holter monitor, respectively. Both measurement systems have been demonstrated to provide sufficient data quality [16, 17]. Notably, the RRITS dataset has been pre-processed and cleaned through the application of a set of criteria defining Holter record validity. In contrast, the authors of the MMASH dataset have not applied any pre-processing techniques to clean the data in their case.

To ensure data integrity and reliability, the raw individual beats undergo meticulous cleaning, following established protocols in the literature [18]. This process implies evaluating non-physiological HR shifts over time to identify and rectify erroneous measurements. Subsequently, beats are aggregated to obtain a more stable granularity. This approach of capturing the average behavior eliminates extreme values and helps to focus on the general trend. A minute-by-minute structure is designed, disregarding data behavior within shorter time frames (less than a minute). The visual inspection of HR processes before and after aggregation and cleaning, as displayed in Fig. 1 for three random users, demonstrates that these transformations result in smoother time series with reduced abrupt changes.

In the RRITS dataset, there are no indications about the specific time of day when data was gathered. Visual inspection suggests that the authors may not have accounted for time spans during which patients were not wearing the Holter monitor. Conversely, the MMASH dataset provides precise time-of-day information, and the percentages of missing values dataset are generally low, especially after aggregation. Specifically, the majority of users consistently wore the sensor, resulting in only a 1% occurrence of missing values. Even in extreme cases, where missing values reached up to 10% over



(a) Before aggregation and cleaning.

(b) After aggregation and cleaning.

Fig. 1 Figures of heart rate time series of 3 random users before and after pre-processing tasks. Before cleaning, several inconsistencies are present which may be caused by motion artifacts. Aggregating helps to smooth heart rate levels whose range was in a non-reliable interval

the course of a full day, this level of data completeness could be deemed acceptable. For such a relatively minor issue, we apply a linear interpolation as an imputation process.

The final pre-processing phase considers stationarity. A process is said to be stationary if all the moments are independent of time series length [19]. Generally, differentiation should be applied to non-stationary time series before processing. Yet, this does not represent a problem since ARIMA can directly handle it. The exploratory phase suggests that the majority of the processes are considered stationary according to the traditional augmented Dickey-Fuller test. To summarize, the limited occurrence of missing values and erroneous measurements underscores the overall high quality of the dataset, offering a solid foundation for our subsequent analysis.

2.3 Adaptive Forecasting Structure

Real-time processing is a critical requirement for modern applications, particularly in fields such as wearable sensor data analysis [20]. These systems must be capable of analyzing continuous data streams as they are generated. This paradigm shift from static to streaming data analysis reflects the growing need for real-world, real-time processing solutions. To address these challenges, the literature highlights adaptive learning strategies for managing real-time data streams, especially in modeling short-term dependencies in HR process. This approach provides effective strategies for handling concept drift, which consists of a gradual shift in data distribution over time [21]. Concept drift is particularly relevant in HR data streams from wearable sensors, where factors such as changes in physical activity, stress, or environmental conditions can change the statistical properties of the data, thus impacting the model's performance. Figure 2 illustrates the schema employed to generate forecasts from the prepared HR data. The forecasting structure is designed for online processing, treating data as a stream rather than in batches. This is crucial in the context of wearable sensor data, which is constantly updated with new observations. Algorithms that do not adhere to online processing may struggle in real-world production environments, as they rely

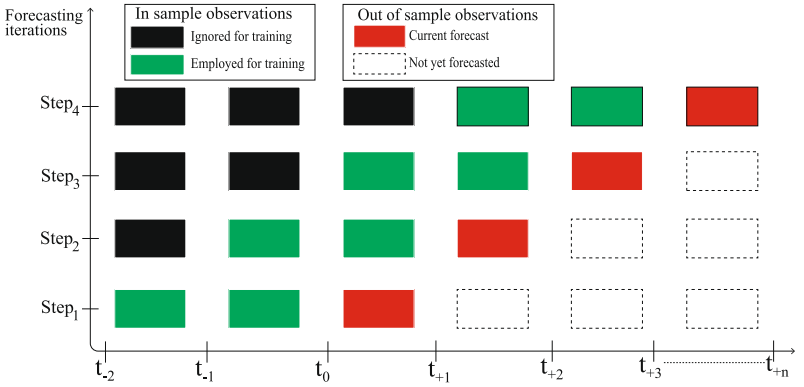


Fig. 2 Schema illustrating the updating procedure. This figure clarifies the data workflow, showing how new observations (green), as they become available, are incorporated into the training process. A rolling window mechanism with a fixed size is employed. At each step, the model forecasts the next observation (red) using a one-step-ahead forecasting mechanism. Black cells represent observations ignored during training, while blank cells are observations that will be forecast in future steps

on historical data that might not reflect current conditions [22]. To address this, we implement two adaptive learning strategies [21].

1. Rolling window. The first strategy involves updating the training data using a rolling window. A fixed size is maintained in the window, which discards the oldest data points while incorporating the most recent ones. By doing so, the model always has access to the latest and most relevant information.
2. Model retraining. Instead of incrementally updating the model parameters, a new model is retrained at each forecasting step. Therefore, the model fully captures the latest data patterns, which may have shifted due to concept drift, rather than relying on previously learned weights. This strategy appears computationally sustainable in the context of short-term HR forecasting and allows the model to better adapt to evolving trends.

To clarify our forecasting strategy, model retraining is conceptually identical to the direct recursive multiple output (DIRMO) strategy [23], although we do not generate multiple outputs simultaneously. Like DIRMO, we retrain the model for each forecasting step, maintaining the model up-to-date.

When dealing with an adaptive learning schema and time series forecasting, researchers should be aware of the type of sliding window they are employing. This parameter can significantly affect the forecast outcome and must be carefully considered [24]. Typically, a sliding window refers to the slice of training data used to make forecasts. In our case, it represents the most recent minutes of HR data, with the number of minutes determined by a parameter we define as the window size. There are two alternative forms of sliding windows: the expanding window and the rolling window [25]. Both methods involve using a dynamic and continuously updating dataset to perform sequential HR forecasts. In an expanding window, there is a fixed origin that retains all past occurrences, whereas a rolling window drops the most distant observations, thus continuously updating the origin. As illustrated in Fig. 2, our approach

employs a rolling window with a constant window size. Therefore, at each forecasting step, a fixed slice of the most recent available records is used for prediction. The key feature of the rolling window method is its adaptability to changing data patterns over time by dropping the oldest observations. This ongoing data refreshment ensures that the forecasting model remains up-to-date and responsive to evolving HR trends.

2.4 Test Validation and Experimental Settings

As shown in Fig. 2, out-of-sample (OOS) validation is adopted to ensure that time dependence is respected [26–28]. OOS observations, held out as the test set, follow chronologically after the in-sample observations used for model training. The OOS set should be sufficiently large to ensure reliable findings, with a minimum of 200 observations recommended [26]. Given our 1-min granularity, this corresponds to an evaluation period of over 3 h. In our study, we set the OOS size to 220 min per user. It is important to note that, with a forecast horizon of one, these 220 observations yield 220 one-step-ahead forecasts for evaluating model performance.

A solid foundation of conclusions, in line with the general principle of cross-validation, is primarily inferred by averaging results when coming from numerous instances. To enhance the number of experiments conducted and, hence, the reliability of our validation procedure, multiple experiments are generated from the combination on various dimensions. Variables that are combined to determine unique experimental settings are

- `Users`: data workflow is replicated for each of the participants in the datasets. Subsequently, resulting scores, employed for comparing specification performance, are the averages of the individual scores within this group.
- `Window size`: the choice of this parameter is crucial when dealing with time series forecasting since it affects resulting scores [24]. A set of values for the window size parameter are tested to evaluate its impact. Diverse values allow for a comprehensive assessment, with results presented for each combination of time of the day and window size.
- `Time of the day`: this parameter dictates the timing of data acquisition. It applies only when this information is available and, thus, just for the MMASH dataset. Each user of the MMASH dataset is evaluated during two distinct 3-h periods, one extracted during nighttime and the other during diurnal activities. Tests of different time of the day are presented separately instead of averaging them.

2.5 Formalization of the Validation Structure

A time series is an ordered collection of data points in a time interval, in which each x_t is an observation at a certain time instant. Let X_t be a time series of length t defined as

$$X_t = [x_1, x_2, \ldots, x_t]. \tag{1}$$

Forecasts generated during validation could be defined by the vector X_N, where N represents the constant number of OOS observations to predict. However, to better represent the forecasting structure adopted to test our algorithm, we prefer the alternative representation $\hat{X}_{t+1}^{(t+N)}$. In this notation, subscript and superscript identify respectively the first and the last element in the sequence. The vector of forecasts generated at time t could be formalized as

$$\hat{X}_{t+1}^{(t+N)} = [\hat{x}_{(t+1)}, \hat{x}_{(t+2)}, \ldots, \hat{x}_{(t+N)}]. \tag{2}$$

where $\hat{X}_{t+1}^{(t+N)}$ represents the entire vector of forecasts for N time steps; $\hat{x}_{(t+1)}$ represents the forecast element generated at time $t+1$, initiating the sequence of forecasts following the present time t. The vector $\hat{X}_{t+1}^{(t+N)}$ is estimated as

$$\hat{X}_{t+1}^{(t+N)} = F_p(X_w). \tag{3}$$

where w ($w \in [1, W]$) specifies the size of the rolling window at each iteration; p refers to the number of iterations necessary to forecast all the N OOS observations. Since the *horizon* $H = 1$, p corresponds to N. Indeed, the number of steps $p = N/H$; F_p represents p different forecasting models employed during validation; X_w is a vector containing w actual observations which are continuously updated to forecast given the last available values.

2.6 ARIMA Algorithm and the Random Walk Model

ARIMA algorithm is adopted to forecast. ARIMA is a family of models renowned for its ability to capture a wide range of complex temporal patterns. Versatility of the method is a remarkable motivation under its recurrent usage and is primarily attributed to its hyper-parameter tuning: p, d, and q. Its general equation in its explicit form could be expressed as

$$(1-L)^d X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \vartheta_j \varepsilon_{t-j} = \ldots \quad . \tag{4}$$

From Eq. 4, components of ARIMA are inspected:

1. Autoregressive (AR) component (p): p in ARIMA reflects the adaptability of the model to different data processes. It signifies the order of autoregressive terms and, consequently, how far back in time the model reaches to understand data dependencies. A higher p accommodates longer-term dependencies, making ARIMA capable of capturing intricate patterns in the data.
2. Differencing (I) component (d): d in ARIMA represents the order of differencing required to achieve stationarity through the usage of lag operator L. This value caters to different levels of data volatility, ensuring that ARIMA can handle both mildly and heavily trending time series.
3. Moving average (MA) component (q): q in ARIMA embodies its adaptability to short-term data dynamics. It signifies the order of moving average terms and,

consequently, the extent to which ARIMA considers recent noise or errors. A higher q allows ARIMA to capture subtle, short-term fluctuations in the data, making it versatile across a spectrum of data complexities.

This adaptability enables ARIMA to capture various data processes. On the one hand, complex time series, presenting intricate and highly volatile data, tend to be better modeled by higher values of p, d, and q. On the other hand, in cases where the data exhibits moderate or mild patterns, a more parsimonious choice of hyper-parameters ensures that model complexity is not needlessly increased, maintaining it particularly effective for capturing simpler processes. A specification refers to a chosen configuration of these hyper-parameters, enabling insight extraction through the assessment of predictive capacity. ARIMA thus emerges as a powerful tool due to its interpretable nature, which facilitates a clear understanding of underlying dynamics.

In this study, addressing our research question involves examining whether simpler models can adequately forecast HR time series. Notably, ARIMA (0,1,0), which corresponds to the random walk model, may serve as a baseline for comparison as it is one of the simplest specification, with only the parameter d beyond the constant term. A random walk model employs only two parameters: the intercept and σ^2. At each minute, a new forecast is generated by randomly adding a value to the last available observation, captured by the parameter intercept. The magnitude of this added value is regulated by the σ^2 parameter, which captures the recent variance of the HR process.

2.7 Residual Diagnostics

Residual diagnostics represents a promising tool, sometimes overlooked in machine learning research which focuses on comparative analysis for assessing predictive performance. Box and Pierce [29] suggested that, after identifying the best specification, residuals should be analyzed to evaluate whether the model still lacks of fit. The Shapiro-Wilk and Kolmogorov-Smirnov tests are adopted to evaluate the normality of residuals [30]. To assess the independence of residuals, autocorrelation is examined, as its presence may suggest that the algorithm is not capturing all explainable variance. The Durbin-Watson test [31] generates a test statistics where values outside the interval (1.5, 2.5) indicate presence of autocorrelation in the residuals. Traditionally, diagnostic methods are used to assess model fit, with normality and independence in residuals being desirable properties. However, in this study, instead of relying solely on statistical tests, we prioritize visual inspection to gain deeper insights into the random walk model. We deliberately opted for the histogram distribution due to the practical difficulty of plotting multiple residual plots simultaneously with the other techniques. Therefore, we examine the distributional patterns of the forecasts generated, focusing on studying the shapes approximating residual distribution. This approach combining visual and statistical diagnostics supports our investigation into the random walk model.

2.8 Evaluation Metrics

The selection of a particular evaluation metric depends mostly on the interpretability and comparability with prior studies. Typically, two of the most widely used evaluation metrics are mean absolute error (MAE) and mean absolute percentage error (MAPE). MAPE, expressed in percentage, offers an intuitively straightforward measure of accuracy. However, its adoption requires caution due to its susceptibility to generating misleading metrics when dealing with actual values ranging in the interval (-1, 1). Indeed, inflated percentage errors arise from the use of these values as denominators. In HR forecasting, this limitation may occur especially if other researchers apply differentiation before processing data. As such, MAPE should be avoided in this context, with MAE representing a more optimal alternative. MAE strength relies on its simplicity, as it reports errors on the same scale or in the same units as the original data. We argued that MAE is preferred over its relative transformation (i.e., root mean square error or mean square error), given the absence of evidence in the literature to support the attribution of different weights in this scenario.

3 Experiments

Autoregressive methods have demonstrated superior performance when addressing HR forecasting issue. However, researchers have not inspected which is the optimal specification of ARIMA. Therefore, a comprehensive forecasting analysis is conducted by comparing performance of various specifications of ARIMA hyper-parameters. We aim at generating an interpretable model in which hyper-parameter combination may provide crucial insights.

3.1 Experimental Setup

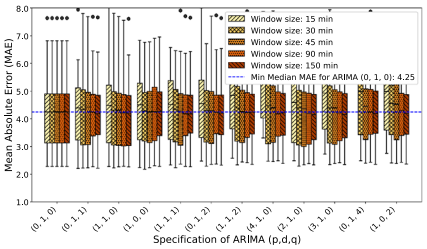
To assess which is the best specification of ARIMA in a minute-by-minute forecast, different combinations are tested. Specifically, a reduced set of candidates has been identified by comparing performance in a step-wise manner. In the optimization process, first tests suggested that the algorithm tends to exhibit inferior performance when specifications contain larger values. Therefore, the following conditions are defined to identify the set of specifications to be tested: *(i)* p lower than 5; *(ii)* d lower than 2; *(iii)* q lower than 5. However, some exceptions with larger values are included in the results to provide insights that reinforce the tendencies mentioned above.

Window sizes are studied by comparing the performance of the algorithm on a limited set of values. The algorithm is tested on the following values: 15 min, 30 min, 45 min, 90 min, and 150 min. Instead of just selecting the optimal value, resulting scores are displayed for each one of the window size value to enable an assessment of algorithm behavior. Tests are not only generated separately with respect to the window sizes, but also for each dataset. Furthermore, with regard to the MMASH dataset, tests are duplicated according to the parameter time of the day, which regulates the selection of the slice of data. The distinction between day and nighttime data is necessary since

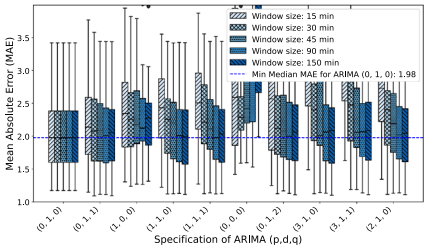
HR properties differ significantly, with higher BPM values typically observed during daytime activities. In our case, MAE, a scale-dependent measure, would have captured more bias from tests performed during the day, averaging larger numerical values [27].

3.2 Forecasting Results of Most Performing Specifications

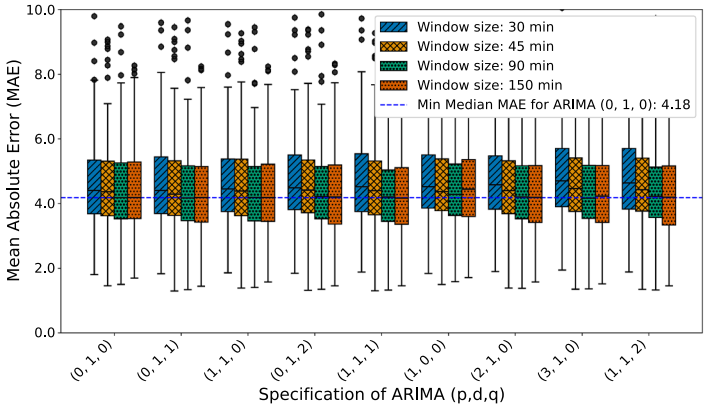
Figure 3 presents the boxplots of MAE scores derived from the time series of the MMASH dataset, showing separate figures for each time period (diurnal and nocturnal), with the RRITS dataset plotted below. Boxplots are used here to analyze the distribution of MAE scores, where each data point corresponds to the average error for an individual user under a specific experimental condition. Therefore, this illustration offers a comprehensive view of variations in the error metric across each setting. Instead of selecting only the top-performing specification, we opted to dis-



(a) Tests on diurnal MMASH time series.



(b) Tests on nocturnal MMASH time series.



(c) Tests on RRITS time series.

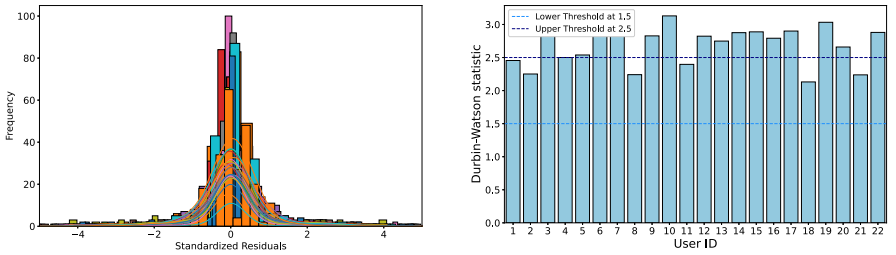
Fig. 3 Each boxplot displays the mean absolute error (MAE) for the top six performing ARIMA specifications. The first two plots present MAE results from the MMASH dataset, with the left plot (in warm colors) for the diurnal time series and the right plot (in cool colors) for the nocturnal time series. Below these, the boxplot from the RRITS dataset is shown. In all figures, each specification is assessed over different window sizes, with the resulting MAE displayed accordingly. A horizontal dashed line indicates the lowest median MAE across all window sizes and specifications, with the relative minimum median value noted in the legend

play the top six performing ARIMA specifications based on their median performance (indicated by the horizontal line within each box). The selection was applied before plotting by separately assessing MAE scores for each dataset and for each window size. Consequently, figures display more than six specifications if the most performing models varied across window sizes. For example, the figure on nocturnal time series for MMASH includes 10 distinct specifications, which differ in part from those selected for the diurnal time series. The following observations can be drawn from the figure:

- Limited exploitation of past information by ARIMA specifications. In the top six performing, the most performing are the simplest, indicating limited reliance on historical data in univariate HR forecasting. Specifically, ARIMA (0,1,0), a random walk model, ranks as a top performer. Similar weak models, such as ARIMA (1,1,1) and (0,1,1), yield comparable results. This trend suggests that little information needs to be leveraged from the inclusion of past values or error terms for improving forecasting accuracy.
- Assessment of window sizes performance. It seems that 90-min and 150-min window sizes perform better overall. Particularly, more complex specifications benefit more from a wider window size as they need more data to train on. Indeed, the simplest model, the random walk, exhibits the same score across all the different window sizes in the MMASH dataset, while slightly differs just in the RRITS dataset. However, while the benefits of larger window sizes are more evident for complex models, as yet stated, forecasting does not need complex models, even if trained on a wide window size.
- Necessity of differentiation. It seems that including one term of differencing, d equal to one, generally enhances forecast accuracy.
- Comparison of diurnal, nocturnal, and RRITS MAE. Diurnal data from MMASH generally result in higher MAE compared to nocturnal data, yet this does not necessarily indicate poorer performance during the day since MAE is scale-dependent, and diurnal BPM tends to be higher. Additionally, the RRITS results closely mirror the MMASH diurnal findings, both in MAE and in the model specifications. This similarity suggests that RRITS may capture daytime activity and exhibit similar HR variability patterns to MMASH diurnal data.

3.3 Random Walk Inspection Through Residual Analysis

Additional insights into forecasting model performance could be elicited from analyzing residuals. Residuals of the random walk model from the 22 MMASH user time series are inspected. P-values from the Shapiro Wilkinson and Kolmogorov-Smirnov tests, collected in Table A1, do not support the statistical normality of residuals. However, for our study, it is more relevant to examine graphically the residuals in detail to better assess the random walk nature and adequacy. Figure 4 displays histograms of standardized residuals. Even if most curves in the figure roughly approximate a normal distribution, the standard deviations appear lower than that of a standard normal distribution. Unlike a standard normal distribution where about 68% of data lies within $\pm$ 1 interval, less variability is evident in the nearly zero, flattened tails. Therefore,



(a) Distribution of residuals (b) Durbin Watson test statistics

Fig. 4 Residual diagnostics of the 22 users in the MMASH dataset. Normality is evaluated from the distribution of residuals. In the y-axis is reported the frequency, whereas in the x-axis is the value of standardized residuals. Durbin Watson statistics are employed to assess the presence of autocorrelation in residuals. Values above the threshold at 2.5 are associated to negative autocorrelation, while below threshold at 1.5 to a positive autocorrelation

residuals are predominantly concentrated around 0, which could be associated to a good performance of the predictor. Lastly, the Durbin Watson test is performed to assess whether autocorrelation is present in the residuals. The test statistics, collected in Table A1 and displayed in Fig. 4, are predominantly near the upper threshold of 2.5, indicating a tendency toward negative autocorrelation.

4 Discussion

In Fig. 5, forecast HR values initially appear to align closely with actual HR values. However, upon closer inspection, the forecast sequence behaves more like a shifted

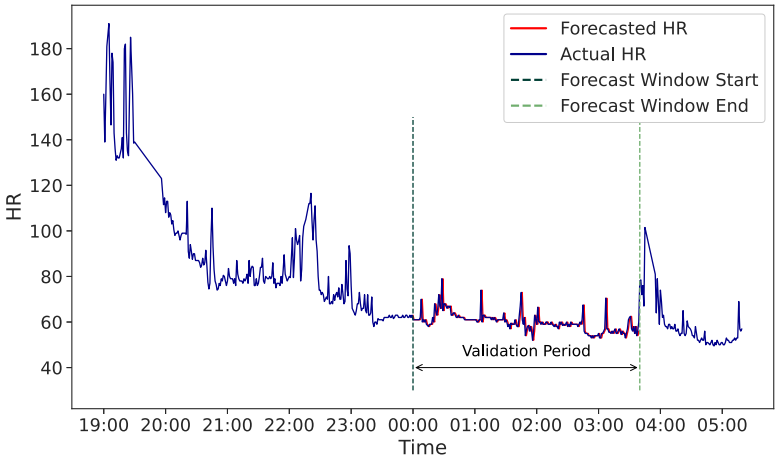


Fig. 5 Forecasts of the random walk model (red) against actual values (blue). Forecasts of one random user are extracted and plotted against the relative actual values to visually inspect the goodness of the prediction. On x-axis, time expressed in hours indicates that forecasts are those generated during nighttime. ARIMA (0,1,0) is employed to carry out these forecasts from a time series of the MMASH dataset

version of the actual sequence. This observation raises a key question about the role of historical HR data in informing future predictions within a univariate and short-term forecasting setting. By evaluating different ARIMA specifications, we aim to determine how past data is leveraged to generate forecasts.

In finance, random walk theory posits that market prices follow an unpredictable path, making historical information minimally informative for future predictions [11]. As Fama noted, the random walk model may serve as a baseline for assessing analyst performance [11]. Similarly, we use ARIMA (0,1,0), which corresponds to the random walk model, as a benchmark to address our research question.

Our findings indicate that, despite the recent trend of using complex machine learning and deep learning models, the random walk model remains competitive, if not superior, as a forecasting baseline. This outcome may be surprising, given that the random walk model only captures stochastic variance around recent values by regulating a single parameter beyond the intercept. For this reason, one might expect the random walk model to often fall short. However, research in financial forecasting has explored why even sophisticated models struggle to outperform the random walk model [32], emphasizing that this challenge is not uncommon. Furthermore, residual diagnostics support this observation by providing insights into the forecast accuracy of the random walk model. While non-normality in residuals is traditionally viewed as a potential limitation, this does not necessarily represent a drawback [33]. In this case, forecast accuracy seems to be acceptable, as a high concentration of residuals around zero corresponds to effective predictor performance. Additionally, the tendency of negative autocorrelation aligns with the random walk process, where values shift randomly from the previous observation.

5 Related Works

The advent of wearable technologies in recent years has fostered further research on HR time series. In the literature, there is a plethora of studies analyzing HR, with several research directions. For instance, some researchers explore the inferential implication of HR variability, particularly its relevance to the activities of the nervous system [34–36]. Another significant research area is the development of early warning systems for detecting cardiac conditions, which is primarily led by ECG-based methods. Today, ECG information can be continuously gathered by wearable sensors. The time intervals between two consecutive heart beats, which we use in our analysis, are embedded within the ECG waves, specifically measuring the elapsing time between two R peaks. This is underscored in our second dataset, “RR interval time series from healthy subjects” (RRITS), which employs a Holter monitoring device to collect ECG data. Most efforts in this domain focus on detecting the presence of arrhythmia, though recent research has shifted towards screening for arrhythmias onset in advance [37, 38].

Although findings from ECG-based studies are promising and often supported by comprehensive methodologies, the reliability of many other current early detection systems remains questionable due to their reliance on cross-sectional data [39–41]. These attempts to develop early detection systems typically use RR intervals and other

physiological parameters, without leveraging full ECG information. The Cleveland dataset or similar surrogates are often used, which only offer a snapshot of clinical conditions. We argue that findings based on such datasets are misleading, as they fail to capture the longitudinal nature of cardiac events, making it challenging to claim that reliable alarm systems can be derived from this data. To our knowledge, the few studies that have developed proper real-time monitoring systems without ECG data focus on specific event onsets, such as heart failure or neonatal sepsis [9, 10].

Finally, we turn our attention to the studies most closely related to our work, which focus on forecasting HR in a univariate fashion to develop detection systems. These studies operate under the assumption that accurate forecasting could enable early detection of pathological conditions. This fosters us to assess the predictive power of models, to determine whether this approach is a viable path forward. Furthermore, it has been highlighted that many machine learning practitioners may not be fully aware of specific guidelines related to time series forecasting [24, 27, 28, 42]. Failure to adhere to these guidelines can compromise the validity of findings. Moreover, insufficiently detailed descriptions of the methodology may hinder meaningful comparative analysis among different studies [27]. When reviewing the methodologies used in these studies, it seems that these issues persist, especially when implemented in an online fashion. For example, Luo and Wu [43] do not mention the use of sliding windows and horizons, a non-trivial shortcoming. Conversely, key details (e.g., window size, the horizons, and data acquisition timing) are provided by Lin et al. [3]. However, it is not explicitly clarified whether ARIMA model weights are retrained at each step. Additionally, it is merely tested a particular specification, ARIMA (2,0,0). Four different deep learning models are compared using simulated HR data from the MIMIC database by Alharbi et al. [4]. The forecasting process is clearly outlined, including the lengths of sliding window and horizons. However, the pre-processing steps involve unnecessary transformations of HR time series. Specifically, statistical standardization is applied, which is typically employed only in multidimensional analysis before clustering or distance-based methods. Moreover, just one singular time series of a patient has been extracted, making questionable the validity of the findings. To forecast HR values, we employ autoregressive models. It is important to note that while machine learning and deep learning models have been explored in other works, they have not shown significant improvements over ARIMA in the univariate case. An extensive list of the most important machine learning as well as deep learning models have been proved to underperform compared to ARIMA [5]. Moreover, long-short term memory (LSTM) performances have been tested, and compared to a simple AR (3) model which fits better than power neural networks, AR is a specific case of ARIMA [6]. The performance of autoregressive models against deep learning ones is tested through a comparative analysis by Staffini et al. [6], specifically involving LSTM networks. The study is designed to generate forecasts as minute-based streams, employing an extended 10-day study to provide a robust foundation. The study just lacks details about the adaptive strategies (e.g., sliding window) which are almost surely implemented but not mentioned. Generally, the granularity in these forecasting works is set to 1 min, offering several benefits as outlined in Section [2.2]. Presumably, a granularity at the inter-beat level, on average less than 1 s, is adopted by Oyeleye et al. [5]. However, this choice hinders comparability with other works using a minute-scale

granularity. Forecasting at the inter-beat level typically results in much lower error rates compared to minute-based forecasts. Moreover, predicting values every second may render the solution unsustainable for the deployment.

When dealing with residual diagnostics, various methods can be employed to assess the normality and independence of residuals. The most renowned graphical techniques include the normal Q-Q plot, the normal P-P plot, residuals vs. fitted values plot, autocorrelation function (ACF) plot, and the histogram distribution [44]. Each of these methods offers unique insights:

- Normal Q-Q plot: this plot compares the quantiles of the residuals with the quantiles of a normal distribution, helping to visually assess deviations from normality.
- Normal P-P plot: similar to the Q-Q plot, the P-P plot compares the cumulative distribution function of the residuals with the expected cumulative distribution of a normal distribution.
- Histogram distribution: this method visualizes the frequency distribution of the residuals and can be used to detect skewness, kurtosis, and other departures from normality.
- Residuals vs. fitted values plot: this plot helps in detecting non-linearity, unequal error variances, and outliers. Residuals should ideally be randomly dispersed around the horizontal axis.
- Autocorrelation function (ACF) plot: autocorrelation of the residuals is displayed at different lags, helping to detect any patterns suggesting non-independence.

6 Conclusion

This study presents a comprehensive evaluation of ARIMA-based univariate methods for real-time, minute-by-minute HR forecasting, designed to align with prevailing state-of-the-art practices. Our findings reveal that the random walk model performs competitively, underscoring a critical insight: refining model complexity offers minimal benefit in univariate and short-term forecasting. This suggests that, in this context, historical HR data alone may lack the informational depth needed for significant predictive improvements.

In light of these findings, it becomes evident that the current focus on comparative analysis with increasingly complex models may be misdirected. The limitation lies not within the algorithms themselves, but within the constraints of the univariate, short-term methodology. As such, meaningful advancements may instead be realized by expanding this approach to incorporate multivariate data, such as additional physiological measures, or by focusing on alternative metrics like HR variability.

Moreover, finance research suggests that the random walk model strong performance in short-term horizons may be due to the limited need for historical data beyond recent observations [32]. In short-term forecasting, the need for additional historical data may be minimal, favoring models that rely primarily on recent observations and stochastic variation. Therefore, a shift toward longer forecasting horizons, even if reducing forecast accuracy, could reveal additional patterns over extended timeframes.

However, this approach would necessitate larger datasets, as the current 24-h length may be insufficient to support training for longer-term patterns effectively.

In summary, pursuing these alternative directions could aid in developing real-time alarm systems capable of meeting clinical demands for timely and reliable cardiovascular monitoring.

Supplementary information

Tables of MAE scores for all users, across each dataset, and window size are also available in our private GitHub repository. For review purposes, a zip file containing these tables has been included in the Supplementary Material.

Appendix A. Test Statistics for Normality and Independence of Residuals

Table A1 Test statistics for the different 22 users

User ID	Shapiro *p*-value	Kolmogorov *p*-value	Durbin Watson statistic
1	0.0001	0.0001	2.4551
2	0.0001	0.0001	2.2509
3	0.0001	0.0001	2.9562
4	0.0001	0.0001	2.5004
5	0.0001	0.0001	2.5385
6	0.0001	0.0001	2.9281
7	0.0001	0.0001	3.0163
8	0.0001	0.0001	2.2408
9	0.0001	0.0001	2.8275
10	0.0001	0.0001	3.1298
11	0.0001	0.0001	2.3966
12	0.0001	0.0001	2.8229
13	0.0001	0.0001	2.7487
14	0.0001	0.0001	2.8758
15	0.0001	0.0001	2.8884
16	0.0001	0.0001	2.7918
17	0.0001	0.0001	2.9005
18	0.0001	0.0001	2.1311
19	0.0001	0.0001	3.0330
20	0.0001	0.0001	2.6592
21	0.0001	0.0001	2.2380
22	0.0001	0.0001	2.8799

Shapiro and Kolmogorov are tests for normality, while Durbin Watson statistics are used to assess autocorrelation of residuals

Author Contributions All authors contributed to the design and writing of the manuscript. The implementation and analysis are performed by G.D., while G.S. reviewed the manuscript.

Funding Open access funding provided by Università degli Studi di Modena e Reggio Emilia within the CRUI-CARE Agreement. The authors did not receive support from any organization for the submitted work.

Availability of Data and Materials The dataset MMASH employed in this study has open access on PhysioNet under the Open Data Commons Open Database v1.0 license at the following web URLs: https://physionet.org/content/mmash/1.0.0/, whereas the dataset RRITS has open access on PhysioNet under the Creative Commons Attribution 4.0 International license at the following web URLs: https://physionet.org/content/rr-interval-healthy-subjects/1.0.0/ .

Declarations

Ethical Approval Ethical approval for the usage of de-identified patients data was obtained from the original dataset curators. The dataset MMASH procured via Physionet, which provides de-identified samples. In accordance with the Helsinki Declaration as revised in 2013, this study was approved by the Ethical Committee of the University of Pisa (N. 0077455/2018).

Consent to Participate Not applicable

Consent for Publication Not applicable

Competing Interests The authors declare no competing interests.

References

1. Nichols M, Townsend N, Scarborough P, Rayner M (2014) Cardiovascular disease in Europe 2014: epidemiological update. European Heart Journal. 35(42):2950–2959. https://doi.org/10.1093/eurheartj/ehu299. https://academic.oup.com/eurheartj/article-pdf/35/42/2950/17354923/ehu299.pdf
2. Jagadeeswari V, Subramaniyaswamy V, Logesh R, Vijayakumar V (2018) A study on medical internet of things and big data in personalized healthcare system. Health Information Science and Systems. 6(1):14. https://doi.org/10.1007/s13755-018-0049-x
3. Lin H, Zhang S, Li Q, Li Y, Li J, Yang Y (2023) A new method for heart rate prediction based on LSTM-BiLSTM-Att. Measurement 207:112384. https://doi.org/10.1016/j.measurement.2022.112384
4. Alharbi A, Alosaimi W, Sahal R, Saleh H (2021) Real-time system prediction for heart rate using deep learning and stream processing platforms. Complexity 2021:5535734. https://doi.org/10.1155/2021/5535734
5. Oyeleye M, Chen T, Titarenko S, Antoniou G (2022) A predictive analysis of heart rates using machine learning techniques. Int J Environ Res Public Health 19(4):2417. https://doi.org/10.3390/ijerph19042417
6. Staffini A, Svensson T, Chung U-I, Svensson AK (2021) Heart rate modeling and prediction using autoregressive models and deep learning. Sensors (Basel). 22(1):34

7. Masum S, Chiverton JP, Liu Y, Vuksanovic B (2019) Investigation of machine learning techniques in forecasting of blood pressure time series data. In: Bramer M, Petridis M (eds) Artificial Intelligence XXXVI. Springer, Cham, pp 269–282
8. Alsheikhy A, Said YF, Shawly T, Lahza H (2023) A model to predict heartbeat rate using deep learning algorithms. Healthcare. 11(3):330. https://doi.org/10.3390/healthcare11030330
9. Stehlik J, Schmalfuss C, Bozkurt B, Nativi-Nicolau J, Wohlfahrt P, Wegerich S, Rose K, Ray R, Schofield R, Deswal A, Sekaric J, Anand S, Richards D, Hanson H, Pipke M, Pham M (2020) Continuous wearable monitoring analytics predict heart failure hospitalization. Circulation: Heart Failure 13(3):006513. https://doi.org/10.1161/CIRCHEARTFAILURE.119.006513
10. Ribeiro M, Castro L, Carrault G, Pladys P, Costa-Santos C, Henriques T (2022) Evolution of heart rate complexity indices in the early detection of neonatal sepsis. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 367–372. https://doi.org/10.1109/EMBC48229.2022.9871274
11. Fama EF (1965) Random walks in stock market prices. Financial Analysts Journal 21(5):55–59. Accessed 2024-10-31
12. Chodrow SE, Jahanian F, Donner M (1991) Run-time monitoring of real-time systems. In: Proceedings Twelfth Real-Time Systems Symposium, pp. 74–75 . IEEE Computer Society
13. Stonebraker M, Çetintemel U, Zdonik S (2005) The 8 requirements of real-time stream processing. ACM SIGMOD Rec 34(4):42–47
14. Rossi A, Da Pozzo E, Menicagli D, Tremolanti C, Priami C, Sîrbu A, Clifton DA, Martini C, Morelli D (2020) A public dataset of 24-h multi-levels psycho-physiological responses in young healthy adults. Data. 5(4):91. https://doi.org/10.3390/data5040091
15. Irurzun IM, Garavaglia L, Defeo MM, Thomas Mailland J (2021) RR interval time series from healthy subjects. PhysioNet
16. Hernández-Vicente A, Hernando D, Marín-Puyalto J, Vicente-Rodríguez G, Garatachea N, Pueyo E, Bailón R (2021) Validity of the polar h7 heart rate sensor for heart rate variability analysis during exercise in different age, body composition and fitness level groups. Sensors. 21(3):902. https://doi.org/10.3390/s21030902
17. Gilgen-Ammann R, Schweizer T, Wyss T (2019) RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. Eur J Appl Physiol 119(7):1525–1532
18. Föll S, Maritsch M, Spinola F, Mishra V, Barata F, Kowatsch T, Fleisch E, Wortmann F (2021) Flirt: a feature generation toolkit for wearable data. Comput Methods Programs Biomed 212:106461. https://doi.org/10.1016/j.cmpb.2021.106461
19. Malamud BD, Turcotte DL (1999) Self-affine time series: measures of weak and strong persistence. Journal of Statistical Planning and Inference. 80(1):173–196. https://doi.org/10.1016/S0378-3758(98)00249-3
20. Bent B, Goldstein BA, Kibbe WA, Dunn JP (2020) Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ digital medicine. 3(1):18
21. Gama JA, Žliobaitundefined I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Comput Surv 46(4):1–37. https://doi.org/10.1145/2523813
22. Purohit M, Svitkina Z, Kumar R (2018) Improving online algorithms via ML predictions. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc., ???
23. An NH, Anh DT (2015) Comparison of strategies for multi-step-ahead prediction of time series using neural network. In: 2015 International Conference on Advanced Computing and Applications (ACOMP), pp. 142–149. https://doi.org/10.1109/ACOMP.2015.24
24. Pungitore S, Subbian V (2023) Assessment of prediction tasks and time window selection in temporal modeling of electronic health record data: a systematic review. Journal of Healthcare Informatics Research. 7(3):313–331. https://doi.org/10.1007/s41666-023-00143-4
25. Pesaran MH, Timmermann A (2002) Market timing and return prediction under model instability. J Empir Financ 9(5):495–510. https://doi.org/10.1016/S0927-5398(02)00007-5
26. Cerqueira V, Torgo L, Mozetič I (2020) Evaluating time series forecasting models: an empirical study on performance estimation methods. Mach Learn 109(11):1997–2028. https://doi.org/10.1007/s10994-020-05910-7
27. Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. Int J Forecast 16(4):437–450. https://doi.org/10.1016/S0169-2070(00)00065-0. The M3- Competition

28. Stein RM (2002) Benchmarking default prediction models: pitfalls and remedies in model validation. Moody's KMV, New York, p 20305
29. Box GEP, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 65(332):1509–1526. https://doi.org/10.1080/01621459.1970.10481180
30. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591–611. Accessed 2023-12-04
31. Durbin J, Watson GS (1950) Testing for serial correlation in least squares regression. I. Biometrika. 37(3–4):409–428. https://doi.org/10.1093/biomet/37.3-4.409. https://academic.oup.com/biomet/article-pdf/37/3-4/409/422190/37-3-4-409.pdf
32. Kilian L, Taylor MP (2003) Why is it so difficult to beat the random walk forecast of exchange rates? J Int Econ 60(1):85–107. https://doi.org/10.1016/S0022-1996(02)00060-0. Emperical Exchange Rate Models
33. Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice, 2nd edn. Otexts, Melbourne, Australia. OTexts.com/fpp2
34. Acharya UR, N K, Sing OW, Ping LY, Chua T, (2004) Heart rate analysis in normal subjects of various age groups. Biomed Eng Online 3:1–8
35. Rajendra Acharya U, Paul Joseph K, Kannathal N, Lim CM, Suri JS (2006) Heart rate variability: a review. Med Biol Eng Compu 44:1031–1051
36. Thayer JF, Åhs F, Fredrikson M, Sollers JJ, Wager TD (2012) A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neuroscience & Biobehavioral Reviews. 36(2):747–756. https://doi.org/10.1016/j.neubiorev.2011.11.009
37. Gilon C, Grégoire J-M, Bersini H (2020) Forecast of paroxysmal atrial fibrillation using a deep neural network. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. https://doi.org/10.1109/IJCNN48605.2020.9207227
38. Rooney SR, Kaufman R, Murugan R, Kashani KB, Pinsky MR, Al-Zaiti S, Dubrawski A, Clermont G, Miller JK (2023) Forecasting imminent atrial fibrillation in long-term electrocardiogram recordings. J Electrocardiol 81:111–116
39. Gavhane A, Kokkula G, Pandya I, Devadkar K (2018) Prediction of heart disease using machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1275–1278 . https://doi.org/10.1109/ICECA.2018.8474922
40. Sujatha P, Mahalakshmi K (2020) Performance evaluation of supervised machine learning algorithms in prediction of heart disease. In: 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1–7 . https://doi.org/10.1109/INOCON50539.2020.9298354
41. Nancy AA, Ravindran D, Raj Vincent PMD, Srinivasan K, Gutierrez Reina D (2022) IoT-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. Electronics 11(15):2292. https://doi.org/10.3390/electronics11152292
42. Bergmeir C, Benítez JM (2012) On the use of cross-validation for time series predictor evaluation. Inf Sci 191:192–213. https://doi.org/10.1016/j.ins.2011.12.028. Data Mining for Software Trustworthiness
43. Luo M, Wu K (2020) Heart rate prediction model based on neural network. IOP Conference Series: Materials Science and Engineering. 715(1):012060. https://doi.org/10.1088/1757-899X/715/1/012060
44. Gupta A, Mishra P, Pandey C, Singh U, Sahu C, Keshri A (2019) Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 22(1):67