Forecasting Of Intensive Care Unit Patient Heart Rate Using Long Short-Term Memory

Firdaus Firdaus1*, Muhammad Fachrurrozi2, Siti Nurmaini1, Bambang Tutuko1, Muhammad Naufal Rachmatullah1, Anissa Darmawahyuni1, Ade Iriani Saptiri1, Anggun Islami1, Masayu Nadila Maharani1, Bayu Wijaya Putra3

1Intelligent System Research Group, Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya
2Department of Informatics Engineering, Faculty of Computer Science, Universitas Sriwijaya
3Department of Information Systems, Faculty of Computer Science, Universitas Sriwijaya
*virdauz@gmail.com

ABSTRACT
Cardiac arrest remains a critical concern in Intensive Care Units (ICUs), with alarmingly low survival rates. Early prediction of cardiac arrest is challenging due to the complexity of patient data and the temporal nature of ICU care. To address this challenge, we explore the use of Deep Learning (DL) models, specifically Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Gated Recurrent Unit (GRU), for forecasting ICU patient heart rates. We utilize a dataset extracted from the MIMIC III database, which poses the typical challenges of irregular time series data and missing values. Our research encompasses a comprehensive methodology, including data preprocessing, model development, and performance evaluation. Data preprocessing involves regularizing and imputing missing values, as well as data normalization. The dataset is partitioned into training, testing, and validation sets to facilitate model training and evaluation. Fine-tuning of hyperparameters is conducted to optimize each DL architecture’s performance. Our results reveal that the GRU architecture consistently outperforms LSTM and BiLSTM in predicting heart rates, achieving the lowest RMSE and MAE values. The findings underscore the potential of DL models, particularly GRU, in enhancing the early detection of cardiac events in ICU patients.

Keywords: Cardiac Arrest, Heart Rate, Vital Sign, MIMIC III, Recurrent Neural Network.

1. INTRODUCTION
Cardiac arrest (CA) is a significant issue in Intensive Care Units (ICUs) today, with low survival rates [1]. Predicting CA is challenging due to its complex data characteristics and dependence on the patient's intensive care time [2]. The survival rate for patients with CA is only about 25%. Studies show that around 59.4% of patients exhibit at least one abnormal sign within 1-4 hours before CA, such as respiratory problems or hemodynamic instability. Preventing CA is highly beneficial in reducing hospital mortality, and clinical studies indicate that early detection of CA can reduce mortality [3]. A low heart rate can increase the risk of death, heart disease, and cardiovascular disease; therefore, monitoring heart rate is crucial in identifying
abnormalities and detecting health problems as early as possible. Rapid technological advancements enable the healthcare sector to combine and analyze vast amounts of health data, reducing risks and making more accurate predictions [4].

Vital sign monitoring is a routine part of patient examination. The five vital signs, including blood pressure, respiratory rate, oxygen saturation, body temperature, and pulse rate, serve as valuable data sources for predicting CA, as they are commonly used in clinical practice to detect patient deterioration [3]. Monitoring data is analyzed to evaluate the prediction of vital signs, determining the direction and magnitude of change over time. Some vital signs are tracked using statistically valid methods, while others rely on repeated surveys with unknown accuracy or expert opinion [5]. Predicting vital signs can help doctors and nurses identify critical changes and prevent the patient's condition from worsening, ultimately reducing mortality rates in patients [6]. Heart rate has become a popular indicator of risk for death after a cardiac event [7]. By predicting heart rate values, we can also predict CA events.

Deep learning (DL) is a modern technique in image processing and data analysis that holds great promise for achieving excellent results. DL can effectively solve complex problems quickly [8]. Currently, DL models are successful in extracting context features from raw text [9]. In recent decades, DL-based models have surpassed classical machine learning models in various text classification tasks [10]. DL techniques have also demonstrated impressive performance in various applications [11].

Long Short-Term Memory (LSTM) is a Recurrent Neural Network (RNN) architecture that can overcome the missing gradient problem of neural networks [12]. Compared to basic RNN, LSTM is more efficient because it can better utilize the long-term dependency between data [13]. A scoring system based on DL using RNN has effectively improved sensitivity, and the use of LSTM can detect patients with cardiac arrest and capture the time dependency of time series data [14]. Bidirectional Long Short-Term Memory is an enhanced version of LSTM [11]. In this study, BiLSTM is used because it can learn data in two directions. Gated Recurrent Unit is another version of LSTM used to improve the performance of LSTM, reduce the number of LSTM parameters, and design less complex structures [11].

In this paper, three DL architectures, namely LSTM, BiLSTM, and GRU, are used to predict heart rate values. After conducting several experiments, the best model for predicting heart rate is determined.

2. METHODS

The steps in this research will be made in the form of a framework so that the stages to be carried out are structured and in accordance with the objectives to be achieved. In this study, the vital signs dataset was extracted from the MIMIC III database. This study aims to build a deep learning model that is formed using the right parameters to get the best results. The assessment parameters used in this study were Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The research stages are divided into several stages, namely data acquisition, data pre-processing, model development, model evaluation and drawing conclusions (figure 1).
The vital sign dataset extracted from the mimic III database is characterized by irregular time series tabular data. In addition to being irregular, the extracted data also contains missing values. This dataset encompasses eight vital signs, including heart rate, systolic blood pressure (sysbp), diastolic blood pressure (diasbp), mean blood pressure (meanbp), respiratory rate (resprate), body temperature (tempc), peripheral oxygen saturation (spo2), and glucose. Each data row is accompanied by the ICU stay of each patient (ICUstay_ID) and the recording time (CHARTTIME). In this study, the vital sign variable used is heart rate.

To construct a model, it is essential to have regular data without missing values. Data regularization is achieved by selecting regular data records (at one-minute intervals) and excluding irregular ones.
For missing value imputation, linear interpolation is employed. Linear interpolation fills in missing values by creating a straight line between two given data points.

Data normalization is also performed using MinMaxScaler (equation 1). Scaling has been shown to impact the quality of the model to be built [15]. Even though values are scaled to a range between 0 and 1, the learning process treats the values in the same way. Therefore, scaling improves the overall quality of the model[16].

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

(1)

The preprocessed data is divided into three parts, with each part allocated as follows; 80% for training, 10% for testing, and 10% for validation. The total amount of data used can be seen in Table 1.

<table>
<thead>
<tr>
<th>Windows Length</th>
<th>Prediction Range</th>
<th>Data</th>
<th>Train</th>
<th>Test</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>60</td>
<td>338635</td>
<td>270908</td>
<td>33864</td>
<td>33863</td>
</tr>
</tbody>
</table>

### 2.2 FORECASTING MODEL DEVELOPMENT

In this research, forecasting models are developed by experimenting with various techniques, including LSTM (Figure 2), BiLSTM (Figure 3), and GRU (Figure 4). Each architecture undergoes fine-tuning, a process used to assess model quality based on the configurations employed. Fine-tuning in this study involves determining the values to build the machine learning model. As a point of comparison, the author will construct a model with 50 epochs for comparison with other fine-tuning approaches. Once the best fine-tuning parameters are identified, the author will proceed with 100 epochs for that specific fine-tuning.

The model tuning for vital sign forecasting is conducted with epochs set at 50, 100, and 200, and batch sizes of 16, 32, 64, and 128. In the case of the LSTM architecture, experiments are conducted, resulting in 12 models. These models are generated using 50, 100, and 200 epochs, combined with batch sizes of 16, 32, 64, and 128.
2.3 PERFORMANCE MEASUREMENT

To measure the performance of the model, two main evaluation methods are Mean Absolute Error (MAE) (equation 2) and Root Mean Square Error (RMSE) (equation 3). These two methods give an idea of how well our model predicts the actual values based on the values predicted by the model.
Forecasting Of Intensive Care Unit Patient Heart Rate Using Long Short Term Memory

\[
RMSE = \sqrt{\frac{1}{n} \sum (y_i - y_i')^2}
\]  
(2)

\[
MAE = \frac{1}{n} \sum |y_i - y_i'|^2
\]  
(3)

MAE measures the absolute difference between the actual value and the predicted value. Mathematically, MAE is calculated by summing the absolute difference between each pair of actual and predicted values, then dividing it by the total number of observations. MAE gives an idea of the average prediction error on the same scale as the actual value, regardless of the direction of the error.

Meanwhile, RMSE measures the square root of the average square difference between the actual and predicted values. RMSE gives an idea of the extent to which the prediction error is spread in the data and gives greater emphasis to large errors. RMSE is often used when we want to give a higher penalty for large prediction errors.

By using MAE and RMSE, we can get a more complete picture of our model's performance. When these two metrics have lower values, it indicates that our model has a smaller error rate and a better ability to predict actual values.

3. RESULTS AND DISCUSSIONS

Table 2 provided contains the results of the heart rate prediction models using LSTM, BiLSTM, and GRU architectures with different combinations of epochs, batch sizes, input lengths, and prediction ranges. The focus of this analysis is to examine the performance of the models and identify any patterns or trends. An example of the comparison between the predicted and actual values from the best model can be seen in Figure 5.

TABEL 2.
Performance Results of Heart Rate Forecasting Model

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Epoch</th>
<th>Batch size</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>50</td>
<td>16</td>
<td>0.02016</td>
<td>0.01454</td>
</tr>
<tr>
<td>LSTM</td>
<td>50</td>
<td>32</td>
<td>0.01671</td>
<td>0.01191</td>
</tr>
<tr>
<td>LSTM</td>
<td>50</td>
<td>64</td>
<td>0.01655</td>
<td>0.01172</td>
</tr>
<tr>
<td>LSTM</td>
<td>50</td>
<td>128</td>
<td>0.01671</td>
<td>0.01171</td>
</tr>
<tr>
<td>LSTM</td>
<td>100</td>
<td>16</td>
<td>0.01704</td>
<td>0.01263</td>
</tr>
<tr>
<td>LSTM</td>
<td>100</td>
<td>32</td>
<td>0.01670</td>
<td>0.01190</td>
</tr>
<tr>
<td>LSTM</td>
<td>100</td>
<td>64</td>
<td>0.01704</td>
<td>0.01239</td>
</tr>
<tr>
<td>LSTM</td>
<td>100</td>
<td>128</td>
<td>0.01699</td>
<td>0.01271</td>
</tr>
<tr>
<td>LSTM</td>
<td>200</td>
<td>16</td>
<td>0.01854</td>
<td>0.01323</td>
</tr>
<tr>
<td>LSTM</td>
<td>200</td>
<td>32</td>
<td>0.01797</td>
<td>0.01299</td>
</tr>
<tr>
<td>LSTM</td>
<td>200</td>
<td>64</td>
<td>0.01890</td>
<td>0.01266</td>
</tr>
<tr>
<td>LSTM</td>
<td>200</td>
<td>128</td>
<td>0.01943</td>
<td>0.01346</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>50</td>
<td>16</td>
<td>0.01653</td>
<td>0.01128</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>50</td>
<td>32</td>
<td>0.01638</td>
<td>0.01124</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>50</td>
<td>64</td>
<td>0.01662</td>
<td>0.01175</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>50</td>
<td>128</td>
<td>0.01713</td>
<td>0.01271</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>100</td>
<td>16</td>
<td>0.01669</td>
<td>0.01201</td>
</tr>
</tbody>
</table>
Furthermore, by analyzing the best parameter combinations, we can identify common patterns that lead to optimal performance. For instance, in some cases, using a smaller batch size and a higher number of epochs resulted in better performance. This indicates that training the model for a longer period with smaller batches can improve its ability to learn the underlying patterns.

Overall, the analysis of the best RMSE and MAE values provides valuable insights into the parameter combinations that yield optimal performance. This information can be used as a guideline for selecting and configuring models in similar situations in the future. An example of the comparison between predicted and actual heart rate values can be observed in Figure 5.

![Figure 5](image_url)
4. CONCLUSIONS

Based on the conducted research on forecasting vital signs in Time Series data of intensive care unit patients using the proposed methods, it can be concluded that among the three methods employed, namely LSTM, BiLSTM, and GRU, GRU yields the best results. GRU achieves the lowest RMSE value of 0.01630 and MAE value of 0.01130, obtained with a batch size of 64 and 100 epochs.

In this study, GRU demonstrates superior performance compared to LSTM and BiLSTM methods. The lower RMSE and MAE values indicate that the vital sign predictions made using GRU exhibit smaller errors compared to the other methods. This suggests that GRU has better capabilities in capturing patterns and relationships within the Time Series data of intensive care unit patients.

However, it is important to note that this conclusion is based on the research conducted with specific data and methods employed. These findings may not be directly generalized to different situations or datasets. Therefore, further research and broader testing are needed to validate the superiority of GRU method in different contexts.

REFERENCES