Predicting Hospital Readmissions by CatBoost to Improve Remote Monitoring Systems

Pietro F Magaldi[1,2], Thaynara S Matos[1,2], Júlia A Ferreira[1,2], Jasmine B Nunes[1,2], Pietro CCO Martins[2], Camila R Moreno[2], Guilherme CM Rabello[2], Ahmad A Almazloum[1,2], Emely P Silva[1], Anderson R Rocha[1]

[1] Recod.ai Lab, Institute of Computing, University of Campinas (UNICAMP), Campinas-SP, Brazil
[2] Instituto do Coracão, HCFMUSP, Faculdade de Medicina, University of São Paulo, São Paulo-SP, Brazil

Abstract

Following recent advances in wearable devices and AI classifier models, a system using the CatBoost classifier model to analyze data provided by Smartwatches and cellular devices through remote monitoring system was proposed, in order to improve the accuracy of making the decision in such systems. The input data for each participant were consisted of the patient's medical history along with the patient's vital signals, and statistical features extracted from the signal time series. Vital signals were collected mainly using smartwatches. The model performed binary classification (N=49) across a dataset split into 3 folds, using cross-validation. The Optuna algorithm was used to optimize the model, which scored (98.72 ± 2.22)% balanced accuracy, (95.24 ± 8.25)% F1-score and with (99.15 ± 1.48)% ROC-AUC. Overall, the system showed promising results towards classifying high/low risk patients, given the low number of samples and high evaluation scores. Possible improvements in the project include a larger dataset and model calibration to enhance the reliability of risk scores.

1. Introduction

Recent advances in wearable technology and artificial intelligence (AI) have significantly transformed the landscape of modern healthcare. Wearable devices, such as smartwatches, enable continuous and accurate monitoring of a wide range of physiological parameters, including heart rate (HR) , oxygen saturation (SpO2), systolic and diastolic blood pressure (SBP, DBP) [1]. Simultaneously, AI algorithms have demonstrated remarkable capabilities in analyzing large volumes of biomedical data, including in the area of cardiology [2–4].

In this same context, current remote monitoring systems present limited applicability and efficiency, especially regarding cardiac postoperative patients [5]. This indicates that current systems fail to fully monitor the patients' medical status, which may lead to post-surgical complications, given that these patients are more susceptible to complications such as atrial fibrillation [6].

With this in mind, the main objective of this study is to improve telemonitoring systems by proposing a solution using a CatBoost classifier model to analyze data provided by Smartwatches through remote monitoring system and predict if the cardiac post-operated patient is at possible risk and in need to be readmitted for health checkups. This will not only explore wearable devices' applicability regarding telemonitoring systems, but also aim to improve the quality and efficiency of post-operative patient care by detecting and aiding patients with complications.

2. Methods

2.1. Input Data

The input data for each sample consists in three parts: the patients' medical history (H), vital measurements (V) and time-series signals (TS). In total, data for 55 patients was collected, but due to inconsistencies in time series signals, 6 patients were removed from the study group, resulting in a total of 49 patients. All medical histories and vital measurements were taken at InCor, HCFMUSP, São Paulo by the medical staff.

2.1.1. Medical History

The medical history contains 4 numerical columns: age, height, weight and BMI; 18 columns indicating whether the patient has had: altered heart rate, cardiac insufficiency, thoracic pain, infarction, arterial insufficiency and others; and 4 categorical columns: sex, ethnicity, type of dyslipidemia and surgery (coronary, valve or aorta).

2.1.2. Vital Measurements

Vital measurements were taken before and after the monitoring period of 30 days using both standard devices and SAMSUNG™ Galaxy Watch5 Smartwatches. The following vital signals were collected: HR, SBP, DBP and SpO2, which amounts to 16 columns of numerical data.

2.1.3. Time Series

The time series contains data for 5 vital signals: SBP, DBP, SpO2, HR by Photoplesthysmogram green sensor, namely HR_PPG, and HR by ECG sensor, namely HR_ECG. These measurements were taken by the patient exclusively using smartwatches during the period of 30 days of monitoring post cardiac surgery soon after full recovery. All data collected during the remote monitoring period were obtained using the Web FAPO-SI^3 platform [1]; an internal customized tool.

In total, 17 features were extracted from each these measurements: statistical features: mean, std, max, min, median, q1, q3, skew, kurtosis, rolling mean mean and others; and Fourier transform features: fft mean, fft std, fft max, fft min. Figure 1 represents more clearly the process of data collection. More information regarding data collection present in [1].

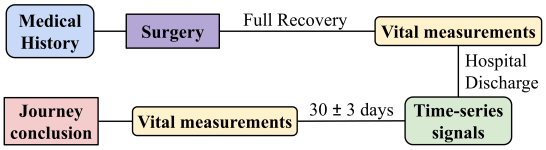


Figure 1. Block diagram representing the process of data collection, with each of the three part that forms the input database.

2.1.4. Patient Labels

Patient labels were extracted as hospital readmissions occurred. Along with readmission labels, additional information was collected, such as readmission date, cause of readmission and doctor observations, as well as the Manchester classification score.

2.2. Preprocessing

Limited preprocessing was performed on the patient's medical history, with a few columns removed in consultation with the medical team, leaving the medical history with 20 columns.

The vital signals measurements had some inconsistencies in its values, mainly missing and NaN values, which were filled with the value zero. It is also important to note that the measurements were not normalized, but instead kept their original value.

The time series required more thorough preprocessing. To ensure temporal consistency, time series values were chronologically corrected. All sequences underwent padding and had their NaN values filled accordingly.

After each component of the input data was preprocessed, they were concatenated and fed to the CatBoost model, as shows Figure 2.

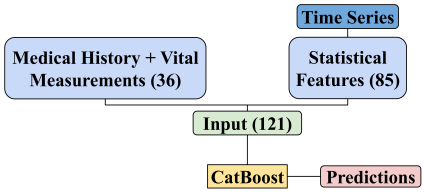


Figure 2. Block diagram for preprocessing and classification, consisted of all part that form the input database.

2.3. Training and Classification

Given the importance of categorical data present in the medical history to the classification, the CatBoost model was chosen. The model and preprocessing parameters were optimized using the Optuna algorithm, and the final model was the result of a study which maximized the F1-score throughout 10000 trials. The model performed binary classification, assigning each sample to one of two classes: readmitted (N=9) or not readmitted (N=40).

3-fold cross-validation was implemented in the experiment to maximize sample usage (N=49). For each fold, the model was trained and tested, and the evaluation metrics were calculated. Figure 3 shows the block diagram of the experiment. The SHAP technique was also used for model explainability.

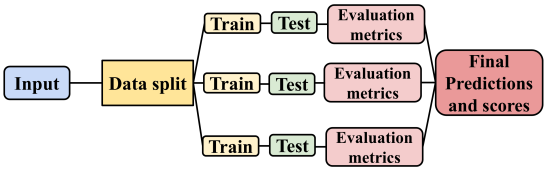


Figure 3. Block diagram of the experiment, representing training, testing and evaluation metrics' extraction.

2.4. Evaluation Metrics

The evaluation metrics were calculated in each fold using reference and predicted labels. For the final scores, the mean was calculated for the 3 folds along with the respective standard deviation. The following metrics were calculated: accuracy, recall, precision, F1, balanced accuracy

and ROC-AUC. Considering the highly unbalanced aspect of the patient labels, maximization of F1 and balanced accuracy was prioritized.

3. Results and Discussion

Overall, the model achieved high evaluation scores, as Table 1 shows. Elevated F1, balanced accuracy and ROC-AUC scores indicate the model was able to correctly classify most of the samples, despite the highly unbalanced classes. Figure 4 demonstrates the ROC curve.

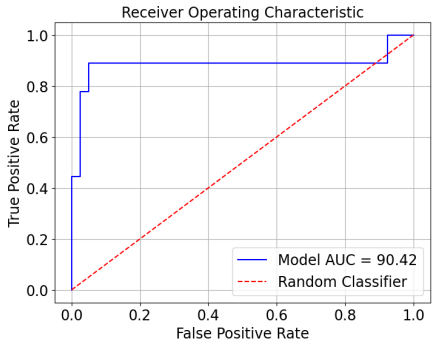


Figure 4. Model's ROC curve, comparing the final model to a random classifier.

In total, there were 2 false positives and 1 false negative, represented in Figure 5. The false negative sample showed symptoms of convulsions, which may indicate the model missed abnormalities in the patient's medical data. Also, the presence of 2 false positives indicate the model may have misinterpreted healthy vital signals.

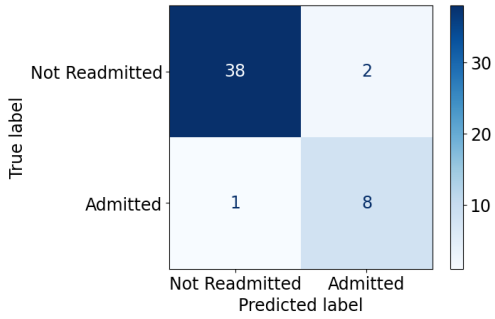


Figure 5. Classification model's final confusion matrix, with 2 false positives and 1 false negative.

When feeding different combinations of the input database to the model, the medical history proved to be the most important feature in sample classification, as its absence in the input dataset resulted in a 50% reduction in F1 score and 28% reduction in balanced accuracy. This is to be expected, given the large clinical and medical value that the medical history carries. Within the medical history, patient weight and age were the most important to sample classification. Also, the inclusion of time-series signals positively affected the results, as Table 2 shows that models trained using the entire dataset scored higher metrics. This indicates that the features extracted from time-series signals positively affected patient classification, which is very promising for the development of remote monitoring using smartwatches.

The role of smartwatch measurements in comparison to standard devices' measurements was also analyzed. Models trained using vital measurements exclusively from standard devices and smartwatches produced nearly identical results, with less than 1% difference in F1 and balanced accuracy scores. This shows that measurements taken from smartwatches are as accurate as those taken from standard devices [1].

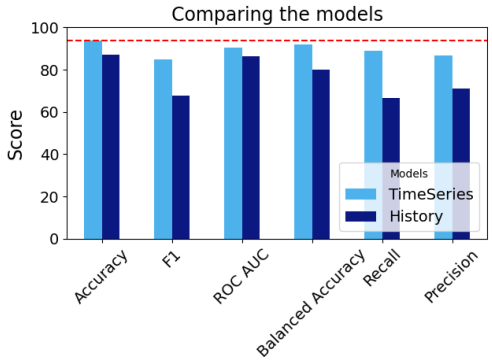


Figure 6. Evaluation metrics' comparative between a model with and without features extracted from time-series.

The results shown in Tables 1 and 2 were obtained with the default classification threshold of 0.5, but it is possible to extract the threshold which maximizes the scores using the ROC Curve. When doing so, a threshold of 0.53 was obtained, which yielded the results shown in Table 3. With the new classification threshold, all results had a significant improvement, as well as the reduction of standard deviation between folds. More importantly, the model was able to correctly classify all the patients who returned to the hospital, as shown in Figure 7.

4. Conclusions

The low number of samples may limit the model's applicability, but overall the model showed promising results towards classifying high/low risk patients. High evaluation metrics: (98.72 ± 2.22)% balanced accuracy, (95.24 ± 8.25)% F1-score and (99.15 ± 1.48)% ROC-AUC indicate that the model was able to efficiently distinguish patients between the two label classes.

Table 1. Evaluation metrics by fold and final scores, with standard deviation as SD.

Fold	Accuracy	F1 Score	ROC AUC	Balanced Accuracy	Recall	Precision
Fold 1	93.87%	80.00%	73.81%	83.33%	66.67%	100.00%
Fold 2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Fold 3	87.50%	75.00%	97.44%	92.31%	92.31%	60.00%
Mean ± SD	93.87 ± 6.25%	85.00 ± 13.23%	90.42 ± 14.44%	91.88 ± 8.34%	88.89 ± 19.25%	86.67 ± 23.09%

Table 2. Evaluation metrics by input, comparing the exclusion and inclusion of time series signals.

Input	Accuracy	F1 Score	ROC AUC	Balanced Accuracy	Recall	Precision
H + V	87.27 ± 6.58%	69.57 ± 18.33%	84.50 ± 11.48%	79.84 ± 14.57%	66.67 ± 28.87%	71.11 ± 7.70%
H + V + TS	93.87 ± 6.25%	85.00 ± 13.23%	90.42 ± 14.44%	91.88 ± 8.34%	88.89 ± 19.25%	86.67 ± 23.09%

H = Clinical History, V = Vital Measurements, TS = Time-series Signals.

Table 3. Evaluation metrics by fold and final scores, using the new classification threshold.

Fold	Accuracy	F1 Score	ROC AUC	Balanced Accuracy	Recall	Precision
Fold 1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Fold 2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Fold 3	93.75%	85.71%	97.44%	96.15%	100.00%	75.00%
Mean ± SD	97.92 ± 3.61%	95.24 ± 8.25%	99.15 ± 1.48%	98.72 ± 2.22%	100.00 ± 0.00%	91.67 ± 14.43%

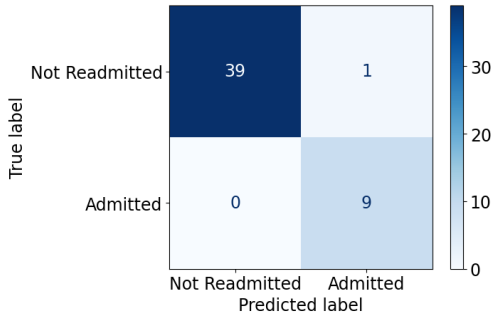


Figure 7. Final confusion matrix with the new classification threshold.

Among all inputs, the medical history proved to be the most impactful in sample classification, with weight and age as the most important columns. Also, the inclusion of data extracted from time series greatly improved the results.

Possible improvements in the project include the use of a larger database and further model calibration to enhance the reliability of risk scores. With appropriate improvements in the proposed model, the resulted applications will not only increase the quality care of the patient, but also reduce the mortality rate coming from postoperative complications.

Acknowledgments

The results presented in this work were funded by Samsung Eletrônica da Amazônia Ltda., under the terms of the Brazilian Informatics Law 8.248/91. The research was approved by IRBCAAE 80963624.2.0000.0068.

References

[1] Monteiro R. Enhancing cardiac postoperative care: a smartwatch-integrated remote telemonitoring platform for health screening with ecg analysis. Frontiers in Cardiovascular Medicine Sep. 2024;11:1443998.

[2] Zhang X. An accurate diagnosis of coronary heart disease by catboost, with easily accessible data. Journal of Physics Conference Series 1955 2021;012027.

[3] Dhananjay B. Analysis and classification of heart rate using catboost feature ranking model. Biomedical Signal Processing and Control Jul. 2021;68:Art. no. 102610.

[4] Wu J. Prediction of three-year all-cause mortality in patients with heart failure and atrial fibrillation using the catboost model. BMC Cardiovascular Disorders Jul. 2025;25(1):466.

[5] Moh'd AF. Postoperative cardiac arrest in cardiac surgery - how to improve the outcome? Medical Archives Apr. 2021; 75(2):149–53.

[6] Smith MEB. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. Annals of the American Thoracic Society Nov. 2014;11:1454–65.

Address for correspondence:

Pietro Fernandes Magaldi
Avenida Albert Einstein, 1251 - Cidade Universitária, Campinas - SP, Brazil, 13083-852
p236842@dac.unicamp.br