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. It scored (91.88 ± 7.40)% balanced accuracy, (83.81 ± 3.30)% F1-score and with (95.18 ± 6.27)% 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 higher number of samples 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, 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, vital measurements and time-series signals. 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 yes/no questions indicating if the patient has had: altered heart rate, cardiac insufficiency and its symptoms, thoracic pain, infarction, arterial insufficiency, cerebral vascular diseases, dementia, chronic lung diseases, connective tissue illnesses, peptic ulcers, liver diseases, diabetes with and without injuries to target organs, hemiplegia, kidney dysfunction, hypertension

and dyslipidemia; 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), which extracted signals from Samsung Health® and Health Monitor® applications and loaded these signals to the database twice a day using a Json extractor.

In total, 17 features were extracted from each these measurements: statistical features: mean, std, max, min, median, q1, q3, skew, kurtosis, rolling mean mean, rolling mean std, rolling std mean, rolling std std; and Fourier transform features: fft mean, fft std, fft max, fft min. The extraction of these features resulted in 85 column of numerical data.

Figure 1 represents more clearly the process of data collection. More information regarding data collection present in [1].

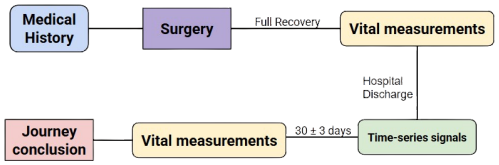

Figure 1. Block diagram representing the process of data collection, with each of the three part that forms the input database.

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, days passed until readmission, cause of readmission and additional observations from the doctor, as well as the Manchester classification score, which indicates the patient's urgency.

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. To minimize the impact of these values in the classification, they 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. The missing and NaN values were filled with a random value from the interval, as to maintain data integrity. The sequences had varying lengths; to standardize them, the mean length was computed and all sequences were padded to match this average length. Padding followed the same approach as used for handling NaNs. Specifically in the HR time series, there were a large quantity of zero values, and they were also filled with random values from the interval.

After each component of the input data was preprocessed, they were concatenated and fed to the CatBoost model. Figure 2 demonstrates more clearly how the input data was obtained.

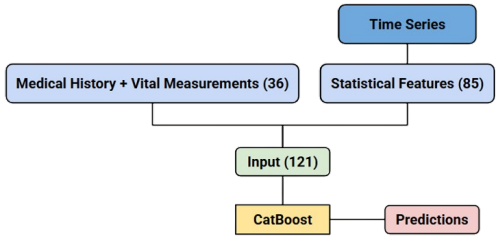

Figure 2. Block diagram for preprocessing and classification, consisted of all part that form the input database.

2.3. Training and Classification

Given the importance of categorical data present in the medical history to the classification, a model that focused on the encoding of these variables was required. CatBoost's ordered target encoding enabled a more detailed and complete representation of categorical data. The model was 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).

Cross-validation was implemented in the experiment to maximize sample usage (N=49). 3-fold stratified cross-validation was chosen for the experiment, in order to maintain, in each fold, the proportions between the two classes. For each fold, the model was trained and tested, and the evaluation metrics were calculated. Figure 3 shows the block diagram of the experiment.

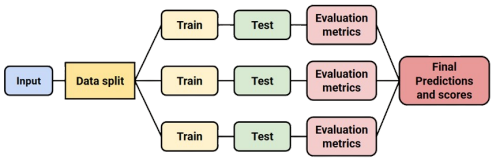

Figure 3. Block diagram of the experiment, representing training, testing and evaluation metrics' extraction.

2.4. Evaluation Metrics

The evaluation metrics were calculated in each fold using reference and predicted labels. For the final scores, the mean was extracted among all folds, along with the standard deviation they presented. 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 and balanced accuracy scores indicate the model was able to correctly classify most of the samples, despite the highly unbalanced classes. The high ROC-AUC demonstrates the model's effectiveness in distinguishing between the different classes. Figure 4 demonstrates True Positive Rate (TPR) and False Positive Rate (FPR) across different classification thresholds, comparing the final model to a random classifier, represented by the red line.

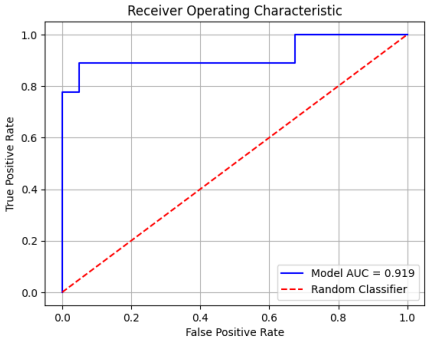

Figure 4. Model's ROC curve, comparing the final model to a random classifier.

In total, there were 2 false positives and 1 false negative, represented in Figure 5. Although information regarding Manchester classification was missing for the false negative sample, the patient showed symptoms of convulsions, which may indicate the model missed some abnormalities in the sample's medical data. Also, the presence of 2 false positives indicate the model may have misinterpreted healthy vital signals, which caused it to misclassify healthy patients.

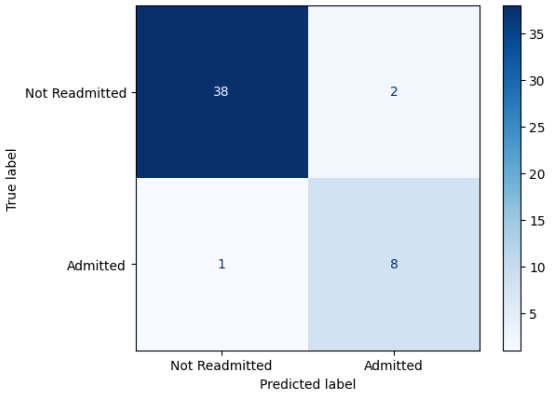

Figure 5. Classification model's final confusion matrix, with 2 false positives and 1 false negative.

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 the largest drop observed in evaluation metrics: 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. 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 help the model classify high/low risk patients as effectively as measurements taken from standard devices.

Another relevant aspect is the impact of time-series signals on patient classification. Two models were compared in order to visualize how these signals affected the final result: the first underwent training using only medical history + vital measurements, and the second was trained using the entire database. Improvements were observed in all evaluation metrics, and the standard deviation between folds was reduced, as shows Table 2. Figure 6 shows how the different models compare and highlights the ROC-AUC score of 95.18% achieved by the final model. This

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	94.12%	80.00%	88.10%	83.33%	66.67%	100.00%
Fold 2	93.75%	85.71%	97.44%	96.15%	100.00%	75.00%
Fold 3	93.75%	85.71%	100.00%	96.15%	100.00%	75.00%
Mean ± SD	93.87 ± 0.21%	83.81 ± 3.30%	95.18 ± 6.27%	91.88 ± 7.40%	88.89 ± 19.25%	83.33 ± 14.43%

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.

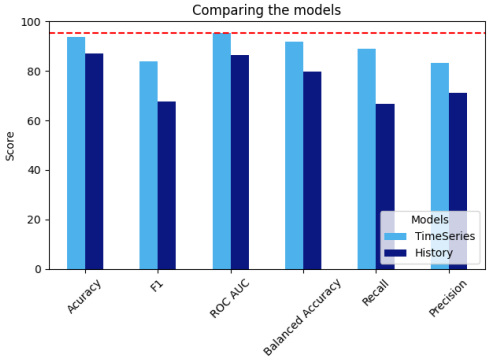

Figure 6. Evaluation metrics' comparative between a model with and without features extracted from time-series.

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, such as (91.88 ± 7.40)% balanced accuracy, (83.81 ± 3.30)% F1-score and (95.18 ± 6.27)% ROC-AUC indicate that the model not only correctly classified the patients' need for hospital readmission, but also that it was able to efficiently distinguish patients between the two label classes.

Possible improvements in the project include the use of a larger database and model calibration to enhance the reliability of risk scores. The inclusion of more patients to the study is in progress and it is expected to provide a more applicable and accurate model. 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 data used in this work were collected by the team from the Instituto do Coração de São Paulo (InCor-SP). 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 IRB-CAAE 80963624.2.0000.0068.

References

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

[2] Zhang X. An accurate diagnosis of coronary heart disease by catboost, with easily accessible data. J Phys Conf Ser 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 Cardiovasc Disord Jul. 2025;25(1):466.

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

[6] Smith MEB. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. Ann Am Thorac Soc 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

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
History + Vital	87.27 ± 6.58%	69.57 ± 18.33%	84.50 ± 11.48%	79.84 ± 14.57%	66.67 ± 28.87%	71.11 ± 7.70%
History + Vital + Time Series	93.87 ± 0.21%	83.81 ± 3.30%	95.18 ± 6.27%	91.88 ± 7.40%	88.89 ± 19.25%	83.33 ± 14.43%