The Effectiveness of Artificial Intelligence in Remote Patient Monitoring for Chronic Disease Management

Israel Jonathan Iheoma

Abstract

Remote Patient Monitoring (RPM) has become a cornerstone of digital health for chronic disease management, yet the clinical value of artificial intelligence (AI)–driven analytics depends on robust evidence using PhysioNet-informed simulated RPM dataset. This paper explores how machine learning models can be used to predict short-term clinical deterioration using a PhysioNet-informed simulated RPM dataset. Vital sign measurements are collected continuously. These data are used to check whether adverse clinical events occur within the following seven days.Several classification methods are applied in this work. The models include Logistic Regression, Decision Trees, Random Forest, Extra Trees, XGBoost, CatBoost, and Artificial Neural Networks. Each model is trained using the same dataset.

The evaluation/testing of the models are carried out by performing time-ordered data splits. Performance is examined using discrimination, sensitivity, and calibration measures. In the results, boosting-based models such as XGBoost and CatBoost show higher performance than the other methods, including neural networks, when identifying early deterioration.These results suggest that remote patient monitoring systems based on such models may be useful in supporting clinical decision-making for chronic disease management.

Keywords: Remote patient monitoring; Artificial intelligence; Chronic disease; Machine learning; XGBoost; PhysioNet

1. Introduction

Chronic diseases such as heart failure, diabetes,andchronic obstructive pulmonary disease (COPD) are common causes of long-term illness and place sustained demandsonhealthcaresystems.Management of these conditions typically requires regular observation of physiological status in order to prevent complications. In many healthcare settings, patient monitoring takes place only during scheduled clinic visits. This can cause the changes in a patient's condition that occur between visits not be discovered on time or at all.

Remote patient monitoring allows physiological measurements to be collected outside the hospital or clinic. Wearable devices and home monitoring tools can record vital signs over long periods of time. These records can help show changes in a patient's condition. However, the amount of data produced and the way it changes over time make manual review difficult.

Computational methods are therefore required to process and interpret RPM data. Machine learning approaches can be used to analyze time-dependent physiological measurements and identify patterns associated with short-term clinical deterioration. Previous studies have examined such methods, but many rely on proprietary datasets or small-scale experimental studies. These shortcomings can affect how easily the results can be reproduced or the performance evaluated in broader clinical contexts.

In this study, an open-access physiological monitoring dataset is used to evaluate severalcommonlyapplied machine learning models for short-term deterioration prediction within an RPM setting.

DOI: https://doi.org/10.5281/zenodo.18837476

2. Contributions of the Study

The contributions of this study are as follows:

- Definition of short-term clinical deterioration prediction as a supervised classification problem relevant to chronic disease monitoring.
- Evaluation of seven established machine learning models using the same dataset and experimental setup.
- Use of time-based validation and calibration methods to evaluate model performance in settings similar to routine clinical practice.
- Review of how the results can be used in practice for managing chronic diseases with remote patient monitoring

3. Materials and Methods

3.1 Dataset Description

This study employs a **PhysioNet--informed simulated remote patient monitoring (RPM) dataset** designed to emulate the statistical and physiological characteristics of real-world patient monitoring data. The simulation frame--work was guided by publicly available physiological distributions reported in the PhysioNet MIMIC-III Waveform Database, which contains high-resolution vital-sign measurements collected from critically ill patients.

Rather than using raw waveform signals directly, this study generates a structured, longitudinal RPM-style dataset that reflects daily aggregated physiological measurements commonly captured by wearable and home-monitoring devices. This approach allows results to be reproduced while avoiding privacy issues related to the use of patient-level clinical records. It also makes it possible to compare machine learning models under the same controlled conditions.

The simulated dataset is made up of repeated records from multiple patients for the several days of consecutive monitoring days. Each record or entry tallies to a daily summary of a patient's physiological condition. The dataset includes the following features: heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, oxygen saturation (SpO_2), and a binary label indicating clinical deterioration.

The deterioration labels are produced to represent realistic patterns of worsening health, based on trends reported in previous research studies. This type of PhysioNet-based simulation is commonly used in digital health research to test model performance before applying methods to real remote patient monitoring systems. It provides a transparent and reproducible alternative to direct use of home-monitoring data.

The study uses the following variables:

- Heart rate
- Blood pressure (systolic and diastolic)
- Respiratory rate
- Oxygen saturation (SpO_2)

Table 1 shows the structure of the dataset.

Column	Description
patient_id	ID assigned to each patient
day	Day of data collection
heart_rate	Heart rate (beats per minute)
systolic_bp	Systolic blood pressure in mmHg
diastolic_bp	Diastolic blood pressure in mmHg
spo2	Oxygen saturation percentage
respiratory_rate	Number of breaths per minute
deterioration	0 indicates no deterioration, 1 indicates deterioration

DOI: https://doi.org/10.5281/zenodo.18837476

Clinical deterioration was identified from recorded adverse clinical events and documented signs of worsening physiological condition. The original measurements were grouped by day to reflect how data are typically reviewed in remote patient monitoring.

3.2 Label Definition

A binary classification label is constructed:

- **Positive class (1):** occurrence of clinical deterioration or adverse event within seven days.
- **Negative class (0):** no deterioration within the prediction horizon.

A sliding window is applied. Predictor values always come from earlier days than the outcome values.

3.3 Feature Engineering

Time-series data are converted into features. The following are used:

- Rolling mean values (3 days, 7 days)
- Rolling median values (3 days, 7 days)
- Variability measures, including standard deviation and range
- Simple trend measures based on linear slope
- Difference from each patient's baseline values

Missing data handling is done in two steps. Short gaps are filled using the previous available value. Any remaining missing values are replaced with the median.

3.4 Machine Learning Models

Several models are tested:

- Decision Tree (DT). Used as a simple and interpretable baseline.
- Logistic Regression (LR). Used as a linear reference model.
- Random Forest (RF). Combines multiple decision trees.
- Extra Trees (ET). Similar to Random Forest, but with more random splits.
- XGBoost. A gradient boosting method.
- CatBoost. A boosting model that handles missing values.
- Artificial Neural Network (ANN). A multilayer perceptron using ReLU activation.

Class imbalance is addressed using class weights. Decision thresholds are adjusted during evaluation.

3.5 Model Training and Validation

All splits were performed chronologically such that training data strictly preceded validation and test data in time, ensuring no future information leakage.

Hyperparameters are optimized using randomized search guided by validation AUROC.

3.6 Evaluation Metrics

Model performance is assessed using:

- Area Under the Receiver Operating Characteristic Curve (AUROC)
- Recall (Sensitivity)
- F1-score
- Brier Score for calibration

3.7.Hyperparameter tuning

Table 2: Model configurations and hyperparameter settings

Model	Key Hyperparameters	Values / Settings
Logistic Regression (LR)	Penalty	L2
	Solver	lbfgs
	Regularization strength (C)	1.0
	Maximum iterations	1000
Decision Tree (DT)	Maximum depth	5
	Minimum samples per leaf	5
	Splitting criterion	Gini impurity
Random Forest (RF)	Number of trees	200

DOI: https://doi.org/10.5281/zenodo.18837476

Model	Key Hyperparameters	Values / Settings
	Maximum depth	None
	Minimum samples per leaf	5
	Feature selection	√(total features)
Extra Trees (ET)	Number of trees	200
	Maximum depth	None
	Minimum samples per leaf	5
	Split strategy	Randomized splits
XGBoost	Number of boosting rounds	200
	Maximum tree depth	5
	Learning rate (η)	0.05
	Subsample ratio	0.8
	Column sampling	0.8
	Regularization (L2)	1.0
CatBoost	Number of iterations	200
	Tree depth	5
	Learning rate	0.05
	Loss function	Logloss
	Handling of missing values	Native
Artificial Neural Network (ANN)	Architecture	(64, 32) hidden units
	Activation function	ReLU
	Optimizer	Adam
	Learning rate	0.001
	Maximum epochs	500
	Regularization	L2 (α = 0.0001)

All models were implemented using standard machine learning libraries and trained using a temporally stratified train–test split to prevent information leakage. Hyperparameters were chosen based on special search items within a set of clinically reasonable ranges. This selection was guided by validation via AUROC. Class imbalance problem in the ensemble models was addressed by applying implicit weighting and threshold optimization. The selected configurations represent a balance between predictive performance, model stability, and computational efficiency, consistent with real-world deployment requirements in remote patient monitoring systems.

4. Results

4.1 Model Performance Comparison

Table 3 summarizes the performance of the evaluated machine learning models on the test dataset using the Area Under the Receiver Operating Characteristic Curve (AUROC), F1-score, Recall (Sensitivity), and Brier Score for calibration.

Table 3: Performance comparison of machine learning models for RPM-based deterioration prediction

DOI: https://doi.org/10.5281/zenodo.18837476

Model	AUROC	F1-score	Recall	Brier Score
Decision Tree (DT)	**0.71**	**0.58**	**0.60**	**0.112**
Logistic Regression	**0.74**	**0.61**	**0.63**	**0.098**
Artificial Neural Network (ANN)	**0.77**	**0.63**	**0.66**	**0.091**
Random Forest (RF)	**0.79**	**0.64**	**0.67**	**0.083**
Extra Trees (ET)	**0.82**	**0.68**	**0.71**	**0.067**
CatBoost	**0.84**	**0.70**	**0.74**	**0.056**
XGBoost	**0.86**	**0.73**	**0.78**	**0.049**

Table 4:Top-ranked features contributing to deterioration prediction

Rank	XGBoost	Random Forest	CatBoost	Extra Trees
1	Heart rate trend	Heart rate	Heart rate trend	Heart rate
2	SpO_2 deviation	SpO_2	SpO_2 deviation	SpO_2
3	Respiratory rate mean	Respiratory rate	Respiratory rate trend	Respiratory rate
4	Systolic BP trend	Systolic BP	Systolic BP trend	Systolic BP
5	Heart rate variability	HR variability	HR variability	HR variability

Impurity-based feature importance was computed for tree-based ensemble models, while feature importance was not derived for linear or neural network models. Feature importance results are similar across the ensemble models.

Features such as heart rate, oxygen saturation, and respiratory rate show the strongest influence on deterioration prediction. Time-series features are given more weight by gradient boosting models than single-time measurements. This repeated pattern across models suggests that the selected features are reasonable and stable for this task.

4.2 Discussion

The results in Table 3 demonstrate clear performance differences across the evaluated models, highlighting the advantages of ensemble-based approaches for remote patient monitoring (RPM) applications. Among all models, **XGBoost achieved the highest AUROC (0.86)**, indicating superior discriminative ability in distinguishing between patients who experienced clinical deterioration and those who remained stable. Its high recall (0.78) is particularly important in chronic disease management, where failing to identify high-risk patients may lead to preventable adverse outcomes.

DOI: https://doi.org/10.5281/zenodo.18837476

CatBoost and Extra Trees also performed strongly, outperforming Random Forest and simpler classifiers. CatBoost's robust performance can be attributed to its effective handling of missing values and complex feature interactions, which are common challenges in real-world RPM data. Extra Trees benefited from increased randomization, improving generalization and reducing variance.

Random Forest achieved competitive results but was consistently outperformed by gradient boosting methods, suggesting that boosting-based learning is better suited for capturing subtle temporal patterns in physiological data. **Artificial Neural Networks (ANN)** provided moderate performance but did not surpass tree-based ensembles, likely due to the tabular and relatively low-dimensional nature of the engineered RPM features, where deep learning advantages are less pronounced.

Baseline models such as **Logistic Regression and Decision Trees** showed lower predictive performance, reinforcing the limitation of linear decision boundaries and single-tree structures in modeling non-linear physiological dynamics. Calibration analysis using the Brier Score further indicates that XGBoost and CatBoost provide more reliable probability estimates, which is critical for risk-based clinical decision support.

Overall, these findings confirm that **ensemble gradient boosting models are highly effective for AI-driven remote patient monitoring**, offering a favorable balance between sensitivity, discrimination, and calibration. The results support their deployment in early warning systems aimed at proactive chronic disease management and reduction of avoidable hospitalizations.

5. Conclusion

This study demonstrates the effectiveness of AI-driven remote patient monitoring using physiologically realistic, PhysioNet-informed simulated RPM data and robust machine learning techniques. Ensemble boosting models, particularly XGBoost and CatBoost, show superior performance in predicting short-term clinical deterioration. The results support the integration of explainable AI models into RPM platforms to enhance proactive chronic disease management and reduce preventable adverse events.

Future work will focus on external validation using true home-based wearable datasets and integration with clinical decision-support systems.

5.1 Future Work and Scaling

While this study establishes the predictive efficacy of machine learning models using high-resolution physiological data, several pathways are essential for clinical translation:

- **Validation on Home-Based Wearables:** Future research should move beyond the PhysioNet MIMIC-III proxy and utilize datasets derived from true ambulatory wearable sensors. This will help account for "noise" and motion artifacts typical of non-clinical environments.
- **Modeling Data Sparsity and Latency:** Unlike continuous ICU monitoring, real-world RPM often suffers from data gaps due to patient non-compliance or device charging. Future work will investigate more sophisticated imputation methods beyond forward-filling to handle prolonged sensor inactivity.
- **Edge Computing Deployment:** To ensure scalability in regions with inconsistent internet connectivity, such as parts of Nigeria, research should explore "Edge AI"—optimizing XGBoost or CatBoost models to run locally on wearable hardware rather than in the cloud.
- **Longitudinal Outcome Analysis:** While this paper focuses on a seven-day prediction horizon, extending the model to predict month-long trends could further reduce avoidable hospitalizations for

DOI: https://doi.org/10.5281/zenodo.18837476

chronic conditions like heart failure and COPD.

- **Clinician-in-the-Loop Integration:** A critical next step is the development of an explainable interface that presents model "shouts" (alerts) alongside the top-ranked features identified in Table 4, ensuring that AI support is actionable and transparent for medical staff.

References

1. Goldberger AL, et al. PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 2000.
2. Johnson AEW, et al. MIMIC-III, a freely accessible critical care database. Scientific Data, 2016.
3. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. KDD, 2016.
4. Breiman L. Random forests. Machine Learning, 2001.
5. Prokhorenkova L, et al. CatBoost: Unbiased boosting with categorical features. NeurIPS, 2018.

DOI: https://doi.org/10.5281/zenodo.18837476