





Predicting High-Risk Hospitalization in Diabetic Outpatients Using Machine Learning: A Real-World Study in a Tertiary Hospital in Brazil

Carolina Deina

Programa de Pós-Graduação em Engenharia de Produção – UFRGS
Av. Osvaldo Aranha, 99 – 5o andar – Porto Alegre, RS
carolinadeina@gmail.com

Flavio Sanson Fogliatto

Programa de Pós-Graduação em Engenharia de Produção - UFRGS
Av. Osvaldo Aranha, 99 – 5o andar – Porto Alegre, RS
flavio.fogliatto@ufrgs.br

Mateus A. dos Reis

Universidade Feevale
Av. Dr. Maurício Cardoso, 510 – Novo Hamburgo, RS
mateusaugustodosreis@gmail.com

Beatriz D. Schaan

Divisão de Endocrinologia – Hospital de Clínicas de Porto Alegre
Rua Ramiro Barcelos, 2350 – Porto Alegre, RS
bschaan@hcpa.edu.br

ABSTRACT

Hospital admissions among diabetic patients pose a major clinical and economic challenge, especially in middle-income countries. This study evaluated the performance of machine learning models in predicting hospitalizations within 12 months, using electronic health record data from 5,964 diabetic outpatients treated at a Brazilian tertiary hospital. Four algorithms were tested: logistic regression, random forest, multilayer perceptron, and gradient boosting. Gradient boosting achieved the best performance (AUC = 0.83; Brier score = 0.165). The most predictive variables included prior hospitalizations, glycated hemoglobin, creatinine, and insulin use. Results suggest that real-world data-driven approaches can help identify high-risk patients and support proactive care strategies in outpatient settings. This study contributes to the development of AI-based clinical tools in the context of Brazilian public healthcare.

KEYWORDS. Diabetes mellitus. Hospitalization prediction. Machine learning.

Paper topics (indicate in order of PRIORITY the paper topic(s)): SA, AS&DS

1. Introduction

Diabetes mellitus (DM) is one of the most prevalent chronic conditions worldwide, associated with high morbidity and frequent hospitalizations. In Brazil, a significant portion of hospital admissions due to DM complications could be avoided with better outpatient care and early detection of deterioration risks [World Health Organization 2016]. Despite advances in health informatics and artificial intelligence, most predictive models are built on data from high-income countries or clinical trials that do not reflect the heterogeneity of public health systems in middle-income countries [Zhang et al. 2010, Miotto et al. 2018]. In this context, predictive models that use routinely collected electronic health records (EHRs) may offer a practical and scalable approach to identifying high-risk patients in advance, allowing for targeted interventions.

In this study, we develop and evaluate machine learning (ML) models capable of predicting the likelihood of hospitalization within 12 months for diabetic patients followed in a large tertiary public hospital in Brazil. Our objective is to assess the feasibility and accuracy of such models when built on real-world outpatient data, and to identify the most informative predictors of hospitalization risk. We also discuss the potential clinical applications and limitations of deploying these models in practice.

2. Methods

We conducted a retrospective cohort study using EHR data from adult patients (≥ 18 years) with a diagnosis of DM treated between 2017 and 2022 in a public tertiary hospital in southern Brazil. The inclusion criteria were at least two outpatient visits during the study period and complete demographic, clinical, and laboratory information. The primary outcome was unplanned hospitalization for any cause within 365 days of an index visit. To reproduce a real-world application scenario, we used a temporal split for model development: data from 2017 to 2021 were used for training and internal validation, while 2022 data served as an external test set. We extracted over 200 variables from structured EHR records, including sociodemographic data, diagnoses (ICD-10 codes), medication prescriptions, lab results, vital signs, and previous utilization (see Table 1).

Table 1 – Summary of baseline demographic and clinical characteristics for the diabetic outpatient cohort

<i>Variable</i>	<i>Value (mean ± SD or %)</i>	<i>Variable</i>	<i>Value (mean ± SD or %)</i>
Number of patients	5,964	Hypertension	78.5%
Age (years)	63.2 ± 11.4	Chronic kidney disease (CKD)	32.1%
Female	57.3%	Insulin use	49.6%
HbA1c (%)	8.4 ± 1.7	Prior hospitalizations (1y)	26.7%
Creatinine (mg/dL)	1.2 ± 0.6	Hospitalizations (12m outcome)	30.9%

Data preprocessing involved imputation of missing values, normalization of numerical variables, and one-hot encoding of categorical features. Five ML algorithms were trained and compared: logistic regression, random forest, XGBoost, support vector machine (SVM), and multilayer perceptron (MLP). Model performance was assessed using the area under the receiver operating characteristic curve (AUC-ROC), precision, recall, F1-score, and calibration. To interpret model outputs, we applied SHAP (SHapley Additive exPlanations) values to identify the most important predictors for individual-level and global explanations [Lundberg and Lee 2017].

3. Results

Among the 5,964 patients analyzed, 1,843 (30.9%) experienced at least one hospitalization within 12 months of a visit. The population was predominantly older (mean age 63

years) and presented a high number of comorbidities. Baseline HbA1c levels were elevated in most patients, and nearly half were using insulin therapy.

The GBM model achieved the best overall performance, with an AUC of 0.83 on the test set, followed by RF (0.80), MLP (0.79), and LR (0.76); see Table 2. Calibration plots indicated that GBM and RF models had the best agreement between predicted probabilities and observed outcomes. Decision curve analysis showed that the GBM model offered a meaningful net benefit across a range of risk thresholds, suggesting clinical utility in flagging patients for preventive interventions.

Table 2 – Comparative performance of different machine learning models on the test set (2022 cohort)

<i>Model</i>	<i>AUC</i>	<i>Brier Score</i>	<i>Sensitivity</i>	<i>Specificity</i>
Logistic Regression	0.76	0.186	0.71	0.68
Random Forest	0.80	0.172	0.74	0.73
Multilayer perceptron	0.79	0.175	0.73	0.72
Gradient Boosting	0.83	0.165	0.78	0.75

Feature importance analysis revealed that prior hospitalizations, number of comorbid conditions, baseline creatinine and HbA1c levels, insulin use, and frequency of outpatient visits were among the most predictive variables; see Figure 1. Notably, time-updated variables and summary statistics over prior months (e.g., lab trends) contributed significantly to model performance, highlighting the value of longitudinal data.

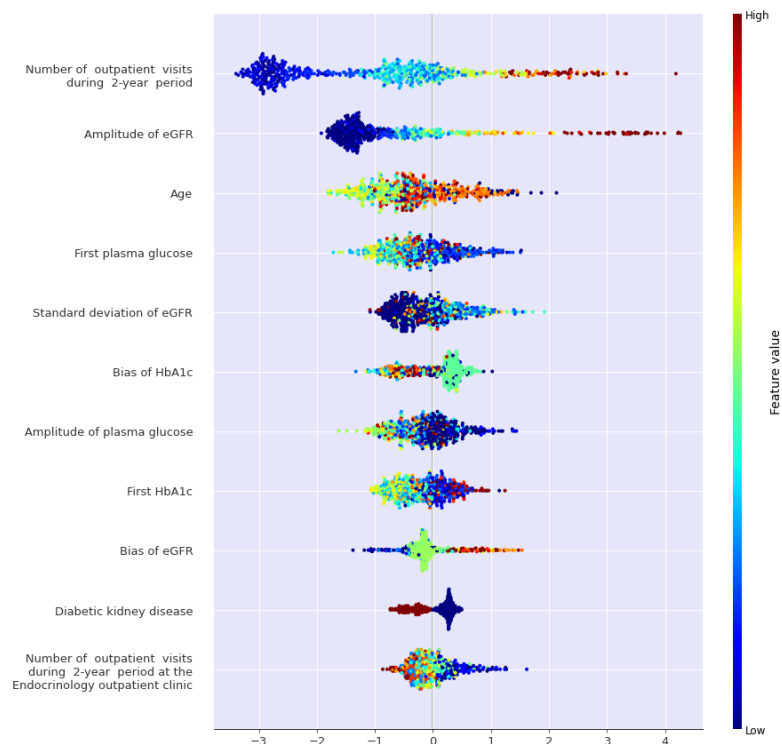


Figure 1. SHAP values (features impacts on model outputs) computed across the entire dataset using the XGBoost-IHT combination model

4. Discussion and conclusion

This study demonstrates the feasibility of using real-world outpatient EHR data to predict hospitalization risk in diabetic patients within the Brazilian public health system. The model achieved moderate-to-high predictive performance, with clinically interpretable results. Key predictors aligned with previous literature on multimorbidity and risk stratification in diabetes [van den Akker et al. 1998, Miotto et al. 2018, Beam and Kohane 2018]. Our findings reinforce the importance of incorporating hospitalization history, comorbidities, and renal and glycemic markers into risk prediction models. Moreover, the use of explainable ML techniques such as SHAP increases model transparency and may enhance clinicians' trust and adoption.

While prior studies have focused on readmission prediction after discharge [Marafino et al. 2018, Rajkomar et al. 2018], our approach targets outpatient populations with no immediate hospitalization event, enabling earlier interventions. Few studies in Latin America have applied predictive modeling in outpatient settings using real-world data [Razavian et al. 2018], highlighting the contribution of this work.

Limitations include reliance on structured data only, lack of information from primary care or private providers, and potential underreporting in administrative records. However, this work provides a step toward deploying ML-based risk stratification tools in resource-constrained settings. Future studies should incorporate temporal modeling and external validation across different healthcare contexts.

References

- Beam, A. L. e Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13):1317–1318.
- Lundberg, S. M. e Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Anais do 31st International Conference on Neural Information Processing Systems (NeurIPS)*, p. 4765–4774.
- Marafino, B. J., Park, M., Davies, J. M., Sowards, S. J. e Boscardin, W. J. (2018). Machine learning for prediction of heart failure readmissions: a systematic review. *PLOS ONE*, 13(11):e0206229.
- Miotto, R., Wang, F., Wang, S., Jiang, X. e Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1:18.
- Razavian, N., Marcus, J., Sontag, D. e et al. (2018). Population-level prediction of type 2 diabetes from claims data and EMRs. *JAMA Network Open*, 1(4):e181676.
- van den Akker, M., Buntinx, F. e Knottnerus, J. A. (1998). Multimorbidity in general practice: prevalence, incidence, and determinants. *Family Practice*, 15(3):189–194.
- World Health Organization. (2016). *Global report on diabetes*. Geneva: WHO.
- Zhang, P., Zhang, X., Brown, J., Vistisen, D., Sicree, R., Shaw, J., et al. (2010). Global healthcare expenditure on diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice*, 87(3):293–301.