Improving Readmission Prediction on Diabetes 130 Dataset

Hailey Reed, Wes Deal, Jared Mercado

Abstract

Hospital readmissions among diabetic patients represent a significant burden on healthcare systems, both in terms of cost and patient outcomes. The Diabetes 130-US Hospitals dataset presents a challenging real-world prediction task due to its high dimensionality, mixed feature types, missing values, and severe class imbalance, with only approximately 11% of encounters resulting in 30-day readmission. While recent work has identified Random Forest–based models combined with synthetic oversampling techniques such as SMOTE as state-of-the-art, the use of artificial medical data raises concerns regarding realism and generalizability. In this project, we investigate imbalance-aware ensemble methods that avoid synthetic data generation. We compare Logistic Regression, standard Random Forest, Weighted Random Forest, and Balanced Random Forest models. Our results demonstrate that Balanced Random Forest improves F1 score by 6.8% over the baseline model and achieves a 4.79% relative improvement over the previously published SMOTE-based Random Forest approach. These findings suggest that balancing strategies embedded directly into ensemble training can outperform synthetic oversampling while remaining grounded in real patient data.

1. Introduction

Diabetes affects over 25.5 million individuals in the United States and is strongly associated with increased rates of hospital readmission. Diabetic patients are nearly twice as likely to be readmitted within 30 days compared to non-diabetic patients, making readmission prediction a critical problem for healthcare systems seeking to reduce costs and improve patient outcomes. Accurate prediction models can enable targeted interventions, better discharge planning, and more efficient allocation of clinical resources.

Despite the availability of large-scale electronic health record data, predicting hospital readmission remains difficult due to heterogeneous patient characteristics, complex nonlinear relationships among clinical variables, and substantial class imbalance. In many hospital datasets, the number of readmitted patients is far smaller than the number of non-readmitted patients, causing standard classifiers to favor the majority class and achieve deceptively high accuracy while performing poorly on the minority class.

This project focuses on improving 30-day readmission prediction for diabetic patients by addressing class imbalance directly within the learning algorithm. Rather than relying on synthetic data generation, we explore ensemble-based methods that rebalance the data during training, with the goal of improving minority-class performance while maintaining realism and interpretability.

2. Dataset

We use the Diabetes 130-US Hospitals for Years 1999–2008 dataset obtained from the UCI Machine Learning Repository (Clore et al., 2014). The dataset contains 101,766 inpatient encounters collected from 130 hospitals across the United States. Each instance represents a hospitalized diabetic

patient who stayed between 1 and 14 days, underwent laboratory testing, and received medication during the encounter.

The dataset includes 47 features, consisting of both categorical and integer variables. These features span multiple domains, including patient demographics, admission type, diagnoses, laboratory procedures, medication usage, and prior hospital utilization. Missing values are present in several fields, reflecting the challenges of real-world clinical data.

The target variable is a binary indicator of whether the patient was readmitted within 30 days of discharge. Only approximately 11% of encounters correspond to positive readmissions, resulting in a highly imbalanced classification problem that strongly influences model performance.

3. Related Works

Recent work evaluating machine learning models on this dataset identified Random Forest–based approaches as the strongest performers. A 2024 comparative study evaluated eleven machine learning algorithms and reported that Random Forest combined with feature selection and SMOTE-based oversampling achieved the best overall performance, with AUC values ranging from 0.74 to 0.78 and F1 scores limited by class imbalance (Liu et al., 2024).

In that study, preprocessing included handling missing values, encoding categorical variables, and converting the target into a binary classification task. Feature engineering involved grouping admission types, creating service utilization variables, and encoding categorical features using one-hot encoding. To address class imbalance, SMOTE was applied during group k-fold cross-validation.

While SMOTE can improve minority-class recall, it generates synthetic medical samples by interpolating between existing patient records. In healthcare settings, this raises concerns about clinical validity and generalization, as synthetic patients may not reflect realistic combinations of diagnoses, treatments, or outcomes. These limitations motivated our exploration of alternative imbalance-handling strategies that rely exclusively on real data.

4. Methodology

4.1 Data Preprocessing

Prior to model training, several preprocessing steps were applied to transform the raw clinical records into a form suitable for machine learning. Missing values were handled using standard imputation strategies, and categorical variables were encoded numerically to allow compatibility with tree-based models. Non-informative identifiers and administrative fields, such as encounter and patient identification numbers, were removed prior to training, as they do not contain predictive information and can introduce spurious correlations.

A key preprocessing step involved the transformation of the three diagnosis fields (diag1, diag2, and diag3). In their raw form, these variables contain high-cardinality ICD-9 diagnosis codes, resulting in a sparse and high-dimensional feature space. Following the methodology proposed by Strack et al. (2014), diagnosis codes were grouped into ten clinically meaningful categories. This grouping reduces

dimensionality while preserving essential diagnostic context, improving model stability and interpretability.

4.2 Models Evaluated

We trained and evaluated four models:

- Logistic Regression as a baseline linear classifier
- Standard Random Forest, representing the core state-of-the-art model
- Weighted Random Forest (WRF), which applies higher misclassification costs to the minority class
- Balanced Random Forest (BRF), which enforces class balance during bootstrap sampling

Logistic Regression provides a point of comparison for more complex ensemble methods, while Random Forest serves as a strong nonlinear baseline well-suited to heterogeneous feature spaces.

4.3 Balanced Random Forest

Balanced Random Forest addresses class imbalance by modifying the bootstrap sampling process used to train each decision tree. Instead of sampling from the original dataset distribution, each tree is trained on a bootstrapped subset containing an equal number of readmitted and non-admitted cases. This forces each tree to learn decision boundaries that treat both classes equally.

Unlike SMOTE, Balanced Random Forest does not generate synthetic samples. All training data correspond to real patient encounters, preserving the clinical realism of the dataset while still mitigating majority-class dominance.

5. Results & Discussion

Model performance was evaluated using the F1 score, which is particularly appropriate for imbalanced classification problems because it balances precision and recall. Accuracy alone is insufficient in this context, as a model predicting all patients as non-readmitted would achieve high accuracy but zero clinical utility.

The Balanced Random Forest achieved a 6.8% improvement in F1 score over the Logistic Regression baseline, demonstrating the importance of ensemble methods for this task. More importantly, when compared to the SMOTE-based Random Forest reported as state-of-the-art in prior work, our Balanced Random Forest achieved a 4.79% relative improvement in F1 score.

These results suggest that explicitly balancing class representation during training is more effective than introducing synthetic minority samples. By maintaining a training process grounded entirely in real patient data, Balanced Random Forest improves minority-class detection while avoiding the potential distortions introduced by oversampling methods.

6. Conclusion

In this project, we examined the problem of predicting 30-day hospital readmissions for diabetic patients using a large, real-world clinical dataset characterized by severe class imbalance. While prior work relied on synthetic oversampling techniques to improve performance, we demonstrated that imbalance-aware ensemble methods can outperform these approaches without generating artificial data.

Our Balanced Random Forest model improved upon both baseline classifiers and the previously published state-of-the-art SMOTE-based Random Forest, achieving higher F1 scores while remaining grounded in real patient encounters. These findings highlight the importance of addressing class imbalance at the algorithmic level, particularly in high-stakes healthcare applications where data realism and interpretability are critical.

Future work could explore combining Balanced Random Forest with domain-informed feature selection or extending these methods to other clinical prediction tasks involving rare but high-impact outcomes.

References

Liu, V. B., Sue, L. Y., & Wu, Y. (2024). *Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes*. *Journal of Medical Artificial Intelligence*, 7, 23. doi:10.21037/jmai-24-70. jmai.amegroups.org

Strack, B., Deshazo, J. P., Gennings, C., Olmo Ortiz, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). *Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records*. *BioMed Research International*, 2014, 781670. doi:10.1155/2014/781670.

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). *Diabetes 130-US Hospitals for Years 1999–2008* [Dataset]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/296/diabetes+130+us+hospitals+for+years+1999+2008