**A Literature Review of Machine Learning Applications Using the Diabetes 130-US Hospitals Dataset**

Prepared by

SHI HEPING

KOH WEI MING, LOUIS

WONG CHEE SENG

National University of Singapore

IT5006 Fundamentals of Data Analytics

Prof Prakash Chandra Sukhwal

Prof Ashish Deepak Dandekar

Updated on September 2025

**Introduction**

The diabetes 130-US hospitals dataset, encompassing clinical records from 130 healthcare facilities across the United States between 1999 and 2008, has emerged as a significant resource for healthcare machine learning research. Originally curated by Strack et al. (2014) (Strack, 2014) to investigate the impact of HbA1c measurements on hospital readmission rates, this dataset contains 101,766 unique patient encounters with over 50 clinical variables (Clore, 2014). The primary objective of most studies utilizing this dataset centers on predicting 30-day hospital readmissions among diabetic patients, a critical healthcare challenge with substantial economic and clinical implications.

**Dataset Characteristics and Clinical Context**

The foundational study by Strack et al. (2014) (Strack, 2014) revealed that HbA1c measurement occurred infrequently (18.4%) in inpatient settings, despite its clinical importance for diabetes management. Their multivariable logistic regression analysis of approximately 70,000 clinical database patient records demonstrated that the relationship between HbA1c measurement and readmission probability depends significantly on primary diagnosis categories. This seminal work established the dataset's utility for examining healthcare quality metrics and predictive modeling applications, showing that greater attention to diabetes reflected in HbA1c determination may improve patient outcomes and lower costs of inpatient care.

The dataset presents inherent challenges typical of healthcare data, including substantial class imbalance with approximately 11% of cases representing readmissions within 30 days. Missing data patterns are significant, with certain variables requiring careful preprocessing strategies across subsequent studies.

**Evolution of Machine Learning Methodologies**

Recent comprehensive studies have demonstrated significant advances in algorithmic approaches. Liu et al. (Liu, 2024) performed an extensive comparison of deep learning and traditional machine learning methods for predicting 30-day readmission rates in patients with diabetes. Their study, utilizing the complete dataset of 101,766 unique encounters from 130 hospitals and integrated delivery networks, found that Random Forest and XGBoost models with Grey Wolf Optimizer outcompeted previous deep learning predictive modeling approaches. The study reported that Random Forest consistently demonstrated superior performance, with the authors noting that machine learning with optimization techniques represents significant progress in predictive analytics for hospital readmission scenarios.

Gandra (Gandra, 2024) included CATBoost in a comprehensive studies of a range of traditional machine learning methods in predicting hospital readmissions among diabetes patients. CATBoost emerged as the top performer thanks to its capability to handle categorical features directly without one-hot encoding and its robustness against imbalanced data.

Shang et al. (2021) (Shang, 2021) conducted extensive evaluations using multiple machine learning classifiers including random forest, Naive Bayes, and decision tree ensemble methods for predicting 30-day hospital readmission risk in diabetic patients. Their analysis of over 100,000 records from the Health Facts Database consistently found random forest algorithms to achieve superior area under the receiver operating characteristic curve (AUC) performance compared to other approaches. The study identified key predictive factors including number of inpatient admissions, age, diagnosis codes, number of emergencies, and sex as critical variables for identifying patients at high risk of short-term readmission.

**Algorithmic Bias and Fairness Considerations**

A critical development in recent literature has been the recognition of algorithmic bias within predictive models derived from this dataset. The open-source Fairlearn project adapted the dataset specifically to examine fairness issues in AI systems, creating a modified version where the target variable "readmitted" was binarized into whether the patient was readmitted within thirty days (Contributors, 2021). This adaptation has been utilized in educational contexts, including the SciPy 2021 tutorial "Fairness in AI Systems: From social context to practice using Fairlearn."

Raza (Raza, 2022) developed a machine learning pipeline specifically designed for predicting, diagnosing, and mitigating health disparities in hospital readmission. This work addresses the critical concern that standard machine learning approaches may result in health disparities caused by biases in data related to social determinants such as race, age, and gender. The proposed methodology analyzes clinical data to determine whether biases exist and removes those biases before making predictions, representing an important advancement in ensuring fairness in healthcare AI applications.

**Technical Challenges and Methodological Considerations**

Class imbalance remains a persistent challenge across studies, with the dataset containing disproportionately fewer positive readmission cases compared to negative cases. Shang et al. (Shang, 2021) addressed this challenge through careful data preprocessing and normalization, ultimately identifying 23 attributes including race, sex, age, admission type, admission location, length of stay, and drug use as modeling risk factors. Their comparison of performance indices across three algorithms revealed that the random forest model demonstrated the best performance with higher AUC than other algorithms. Gandra (Gandra, 2024) handled the issue by employing CATBoost model which mitigate the imbalance using regularization techniques.

Feature engineering approaches have evolved considerably, with studies emphasizing the importance of diagnostic codes and medication patterns. Shang et al. (Shang, 2021) noted that secondary diagnoses (diag\_2) were more important than primary diagnoses among the three diagnostic codes, indicating that subsequent diagnoses in a patient's electronic health record could more accurately reflect the patient's condition and readmission risk.

**Limitations and Future Directions**

Several limitations constrain the generalizability of findings from this dataset. The temporal scope (1999-2008) raises questions about relevance to contemporary healthcare practices, given significant changes in diabetes management protocols, electronic health record systems, and patient populations over the intervening years. The original study by Strack et al. (Strack, 2014) acknowledged that the analysis was undertaken to examine historical patterns of diabetes care and to inform future directions for improvements in patient safety.

Despite achieving high performance metrics in controlled settings, translation from research models to clinical practice remains limited. Few studies report successful real-world deployment of these predictive systems, highlighting the persistent gap between research achievements and practical healthcare applications.

Future research should prioritize development of updated datasets reflecting current clinical practices, enhanced bias mitigation techniques, and rigorous evaluation of model fairness across demographic subgroups. The work by Raza (Raza, 2022) represents an important step toward addressing bias concerns, but broader implementation of fairness-aware machine learning approaches is needed across the field.

**Conclusion**

The diabetes 130-US hospitals dataset has catalyzed significant advances in healthcare predictive modeling over the past decade. The consistent findings across verified studies demonstrate that Random Forest and XGBoost algorithms achieve superior performance for readmission prediction tasks, with recent optimized approaches reaching clinically relevant accuracy levels. However, emerging concerns regarding algorithmic bias, as highlighted by the Fairlearn project adaptations and bias mitigation research, necessitate careful consideration of ethical implications alongside technical performance metrics.

While the dataset has proven valuable for advancing machine learning methodologies in healthcare, the field must address fundamental questions of bias, generalizability, and clinical implementation to realize the full potential of these predictive approaches. The evolution from Strack et al.'s original clinical investigation to current sophisticated machine learning applications demonstrates both the dataset's enduring value and the critical need for continued methodological advancement in healthcare AI applications.

References

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J

Fairlearn Contributors. (2021). Diabetes 130-Hospitals Dataset Documentation. Fairlearn Official Documentation. https://fairlearn.org/main/user\_guide/datasets/diabetes\_hospital\_data.html

Gandra, A. (2024). Predicting hospital readmissions in diabetes patients: A comparative study of machine learning models. International Journal of Health Sciences, 8(3), 289–297. https://doi.org/10.53730/ijhs.v8n3.15189

Shang, Y., Jiang, K., Wang, L. et al. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. BMC Medical Informatics and Decision Making, 21(1), 1-16. https://doi.org/10.1186/s12911-021-01423-y

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. https://doi.org/10.21037/jmai-24-70

Raza, S. (2022). A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission. Healthcare Analytics, 2, 100073. https://www.sciencedirect.com/science/article/pii/S2772442522000430

Strack, B., DeShazo, J. P., Gennings, C., Olmo, 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, Article ID 781670. https://doi.org/10.1155/2014/781670