Prediction of Geriatric Patients Readmission Using XGBoost

Aisha Al-Mustapha[1]; Prema Kirubakaran[2];
Dr. Ridwan Koladapo[3]; Kosiso Precious Nwajiobi[4]

[1,2,3,4] Department of Information Technology, Faculty of Computing
Nile University of Nigeria, Abuja, Nigeria

Publication Date: 2026/08/13

Abstract: Hospital readmission among geriatric patients is a critical indicator of healthcare quality, imposing significant burdens on both patients and healthcare systems. This study proposes an Extreme Gradient Boosting (XGBoost)-based predictive framework for identifying 30-day hospital readmission risk in patients aged 65 years and above. A publicly available synthetic Hospital Readmissions Prediction dataset from Kaggle (n = 30,000) was filtered to obtain 10,645 geriatric records. Following systematic data preprocessing—including blood pressure parsing, categorical encoding, and feature normalization—an XGBoost classifier was trained and compared against a logistic regression baseline. The XGBoost model achieved an overall accuracy of 68.85% versus 63.68% for logistic regression. However, logistic regression demonstrated superior recall (0.4923 vs. 0.3557), indicating greater sensitivity to readmission cases in the minority class. Feature importance analysis identified discharge destination (0.1771), hypertension (0.1082), and diabetes (0.1059) as the most influential predictors. These findings underscore the potential of ensemblebased machine learning to support geriatric care planning, while highlighting the need for improved class-imbalance handling through threshold calibration and resampling strategies. The proposed framework provides a replicable, interpretable decisionsupport tool suitable for integration into electronic health record systems in resource-constrained healthcare environments.

Keywords: *Hospital Readmission, Geriatric Patients, XGBoost, Machine Learning, Predictive Modelling, Class Imbalance, Discharge Destination, Electronic Health Records.*

How to Cite: Aisha Al-Mustapha; Prema Kirubakaran; Dr. Ridwan Koladapo; Kosiso Precious Nwajiobi (2026) Prediction of Geriatric Patients Readmission Using XGBoost. *International Journal of Innovative Science and Research Technology*, 11(8), 254-259. https://doi.org/10.38124/ijisrt/26aug080

I. INTRODUCTION

Population ageing has emerged as a major global public health phenomenon, with older adults constituting an increasingly significant proportion of hospital admissions and healthcare utilization worldwide. Geriatric patients, conventionally defined as individuals aged 60–65 years and above, typically present with complex clinical profiles characterized by multimorbidity, frailty, polypharmacy, and diminished physiological reserve [1], [2]. These factors substantially increase their vulnerability to adverse clinical outcomes, prolonged hospitalizations, and unplanned readmissions, placing considerable pressure on healthcare systems— particularly in low- and middle-income countries where specialized geriatric services are scarce.

In Nigeria, the challenges of geriatric healthcare are compounded by fragile health infrastructure, late hospital presentations, and limited post-discharge continuity of care. High comorbidity burdens—including hypertension, diabetes mellitus, cardiovascular disease, and chronic kidney disease—are frequently documented in elderly Nigerian patients, often coexisting and contributing to repeated hospitalizations [3], [4]. These conditions, when inadequately managed, escalate healthcare costs and strain already overstretched facilities.

Hospital readmission within 30 days of discharge is a widely recognized indicator of care quality and transitional care effectiveness, especially among older adults who are disproportionately vulnerable to post-discharge complications. Traditional approaches to readmission prediction—primarily logistic regression and clinician judgment—exhibit well-documented limitations in modeling the complex, nonlinear

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug080

variable interactions that typify geriatric populations [5]. Machine learning (ML) methods, particularly tree-based ensemble models such as Extreme Gradient Boosting (XGBoost), have demonstrated superior performance in clinical risk stratification tasks due to their robustness to missing values, inherent feature selection, and strong generalization capabilities [6], [7].

Despite encouraging progress in high-income settings, the application of such models within sub-Saharan Africa—specifically Nigeria—remains limited. There is an absence of locally trained, context-specific predictive tools tailored to geriatric readmission risk. This study addresses this gap by developing, training, and evaluating an XGBoost-based predictive model using a structured healthcare dataset filtered for geriatric patients, with the objective of supporting early identification of high-risk individuals and enhancing clinical decision-making.

The remainder of this paper is organized as follows: Section II reviews related work; Section III details the proposed methodology; Section IV presents results and discussion; Section V concludes with contributions and future directions.

II. RELATED WORK

Predicting hospital readmission in geriatric populations has been approached from both clinical and computational perspectives.

Early research predominantly employed logistic regression and survival analysis, offering interpretability but limited capacity to model high-dimensional, nonlinear relationships inherent in electronic health record (EHR) data [8], [9]. Recent studies have consistently demonstrated that ML-based approaches outperform conventional statistical methods for readmission prediction across diverse clinical settings.

Chakraborty et al. [8] demonstrated that applying ML to EHR data enables pre-discharge identification of high-risk patients; however, their study was limited to general adult populations, reducing its direct applicability to the elderly. Abiodun et al. [3] identified cardiovascular disease, diabetes, and chronic kidney disease as primary drivers of geriatric readmissions in African settings but stopped short of developing a predictive model. Ahmed et al. [10] built an XGBoost pipeline with SHAP explainability achieving AUC = 0.81 for adverse outcomes in older adults, though external validation was absent. Okafor et al. [11] applied XGBoost to Nigerian adult hospital data achieving 78.4% accuracy, but geriatric patients were under-represented.

Ensemble methods, particularly gradient boosting variants, have consistently outperformed alternatives in meta-analyses and systematic reviews [12], [13]. González-Nóvoa et al. [14] demonstrated optimized XGBoost for ICU readmission prediction through hyperparameter tuning. Askar et al. [13] highlighted prevalent methodological weaknesses including data leakage, inconsistent validation, and limited external generalizability in the existing readmission ML literature. Longitudinal feature engineering [15] and transformer-based architectures integrating unstructured clinical notes [16] have further improved prediction accuracy, albeit at substantial computational cost.

A consistent gap in the literature is the scarcity of geriatric-specific predictive models validated within resource-constrained, sub-Saharan African healthcare contexts. Most existing frameworks are developed on benchmark datasets such as MIMIC-IV [17], which inadequately represent the Nigerian or broader African clinical environment. The present study directly addresses this gap by targeting the geriatric subpopulation and focusing on clinically relevant, low-complexity feature sets suitable for deployment in resource-limited settings.

III. METHODOLOGY

A. Research Design and Framework

This study employed a quantitative, descriptive predictive research design. The analytical framework was structured around four sequential phases: (1) data acquisition and preprocessing; (2) feature engineering and selection; (3) model development and training; and (4) performance evaluation. The framework operationalizes readmission risk as a binary classification problem— readmitted within 30 days (1) versus not readmitted (0)—and evaluates both a baseline logistic regression model and the proposed XGBoost model.

B. Dataset and Geriatric Cohort Extraction

The Hospital Readmissions Prediction dataset, a synthetic dataset released under the CC0 public domain license on Kaggle, was utilized. It comprises 30,000 patient records with 11 clinical and demographic variables including age, gender, blood pressure, cholesterol, BMI, diabetes, hypertension, medication count, length of stay, discharge destination, and 30-day readmission status. The dataset was filtered to retain only patients aged 65 years and above, yielding a geriatric cohort of N = 10,645 records for all subsequent analyses.

C. Preprocessing

Data preprocessing comprised: (1) removal of duplicate records; (2) parsing of blood pressure strings into separate systolic and diastolic numeric features; (3) binary encoding of categorical clinical variables (diabetes, hypertension, readmission status: Yes/No

→ 1/0); (4) label encoding of nominal variables (gender, discharge destination); and (5) median imputation for any

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug080

residual missing values in numeric columns. Continuous features (age, cholesterol, BMI, systolic/diastolic BP, medication count, length of stay) were retained in their original numeric scales without additional normalization, as XGBoost is scale-invariant. The final preprocessed dataset was confirmed to contain no missing values prior to model development.

D. Feature Engineering and Selection

The decomposition of blood pressure into systolic and diastolic components constituted the primary feature engineering step, increasing the clinical information density of the dataset. Feature selection relied on correlation analysis to identify and exclude highly collinear predictors, supplemented by XGBoost's intrinsic feature importance ranking (gain-based). The final feature set comprised 11 variables: age, gender, systolic BP, diastolic BP, cholesterol, BMI, diabetes, hypertension, medication count, length of stay, and discharge destination.

E. Model Development

Two models were developed and compared. The baseline model was a logistic regression classifier with class_weight='balanced' to account for the class imbalance inherent in the dataset (approximately 12% readmitted cases). The proposed model was an XGBoost classifier (XGBClassifier) configured with n_estimators=150, max_depth=4, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, and a scale_pos_weight parameter computed as the ratio of negative to positive class instances (scale_pos_weight ≈ 7.23). Both models were implemented in Python using scikit-learn and the xgboost library within Google Colab.

F. Training and Evaluation Protocol

The geriatric cohort was partitioned using stratified random sampling into a training set (70%) and a holdout test set (30%), preserving the class distribution across splits. K-fold cross-validation (k=5) was applied on the training set for hyperparameter validation and stability assessment. Performance was evaluated on the independent test set using accuracy, precision, recall, F1score, and ROC-AUC. The area under the receiver operating characteristic curve (ROC-AUC) served as the primary discriminative metric given the class imbalance, supplemented by recall as the clinically critical metric for identifying true readmission cases.

IV. RESULTS AND DISCUSSION

A. Descriptive Analysis of Geriatric Cohort

Table I presents the summary statistics of the study variables across the 10,645 geriatric patient records. The mean patient age was 77.42 years (SD = 7.51; range: 65–90), representing a broad spectrum of younger-old and older-old patients. Mean cholesterol was 224.75 mg/dL, mean BMI was 29.07 kg/m², mean systolic BP was 134.76 mmHg, and mean diastolic BP was 84.98 mmHg.

Patients were prescribed an average of 5.02 medications (SD = 3.18), and mean length of stay was 5.49 days (SD = 2.87). The binary readmission outcome variable had a mean of 0.12 (i.e., 12% of patients were readmitted within 30 days), confirming a pronounced class imbalance that necessitates metrics beyond accuracy for rigorous model assessment.

Table 1: Summary Statistics of Key Study Variables (N = 10,645)

Variable	Mean	Std Dev	Min	25%	50%	75%	Max
Age (years)	77.42	7.51	65	71	77	84	90
Cholesterol (mg/dL)	224.75	43.17	150	188	225	261	300
BMI (kg/m²)	29.07	6.34	18	23.6	29	34.6	40
Systolic BP (mmHg)	134.76	14.62	110	122	135	147	160
Diastolic BP (mmHg)	84.98	8.91	70	77	85	93	100
Medication Count	5.02	3.18	0	2	5	8	10
Length of Stay (days)	5.49	2.87	1	3	5	8	10
Readmitted (0/1)	0.12	0.33	0	0	0	0	1

B. Model Performance Comparison

Table 2 reports the performance metrics for both the logistic regression baseline and the proposed XGBoost model on the independent test set. XGBoost achieved higher overall accuracy (0.6885 vs. 0.6368), reflecting superior classification performance across both classes. However, logistic regression demonstrated considerably higher recall for the positive (readmitted) class (0.4923 vs. 0.3557) and a higher F1-score (0.2477 vs. 0.2172), indicating superior sensitivity in identifying true readmission cases. ROCAUC values were comparable (LR: 0.5822; XGB: 0.5431), with logistic regression holding a marginal advantage in overall discriminative ability.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug080

Table 2: Comparative Performance Metrics of Predictive Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.6368	0.1655	0.4923	0.2477	0.5822
XGBoost (Proposed)	0.6885	0.1563	0.3557	0.2172	0.5431

This performance dichotomy is attributable to the differential optimization objectives of the two algorithms. Logistic regression with balanced class weights penalizes minority-class misclassifications more uniformly, thereby improving recall at the cost of overall accuracy. XGBoost, through its gradient-boosted tree construction, optimizes global log-loss and tends to preferentially classify the majority (non-readmitted) class when minority-class signals are distributed across multiple feature subspaces, as observed in this geriatric dataset.

C. Feature Importance Analysis

Table 3 presents the XGBoost feature importance scores (gain-based). Discharge destination emerged as the dominant predictor (importance = 0.1771), substantially outweighing all physiological variables. This finding underscores that the post-discharge care environment—specifically whether a patient is discharged to a nursing facility versus home—is a stronger determinant of short-term readmission risk than any single physiological measurement. Hypertension (0.1082) and diabetes (0.1059) ranked second and third respectively, reflecting the significant role of chronic comorbidity burden in readmission dynamics. Physiological parameters (cholesterol, BMI, systolic and diastolic BP) and demographic variables (age, gender) contributed moderate and relatively uniform importance scores (0.076–0.078), suggesting that no single physiological measure dominates risk prediction in this cohort. Medication count showed the lowest importance (0.0709), likely reflecting the synthetic and uniform distribution of medication data in the dataset, which attenuates its clinically expected predictive signal.

Table 3: XGBoost Feature Importance Rankings

Rank	Feature	Importance Score
1	Discharge Destination	0.1771
2	Hypertension	0.1082
3	Diabetes	0.1059
4	Cholesterol	0.0783
5	BMI	0.0783
6	Systolic BP	0.0778
7	Length of Stay	0.0761
8	Age	0.0760
9	Gender	0.0759
10	Diastolic BP	0.0756
11	Medication Count	0.0709

D. Benchmarking Against Prior Studies

Table 4 contextualizes the current model's performance within the broader readmission prediction literature.

Table 4: Benchmarking Against Prior Machine Learning Readmission Studies

Study	Algorithm	Dataset	Performance
Okafor et al. [11]	XGBoost	Nigerian adults (hospital)	Accuracy = 78.4%
Ahmed et al. [10]	XGBoost + SHAP	Single hospital dataset	AUC = 0.81
Chakraborty et al. [8]	SVM	Electronic health records	Accuracy = 72%
Li et al. [18]	Ensemble + NLP	ICU EHR dataset	Accuracy = 79%
This Study	XGBoost (proposed)	Kaggle geriatric (10,645)	Accuracy = 68.85%; Recall = 0.36

The lower absolute accuracy of this study relative to prior work is attributable to the pronounced class imbalance (12% readmission rate) and the absence of complex temporal or unstructured features. However, the study's contribution lies in its explicit geriatric focus, transparent methodology, and reproducibility without high-end computational infrastructure.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug080

E. Discussion

The primacy of discharge destination as the leading predictor aligns with established clinical evidence that post-discharge care transitions are central determinants of readmission in older adults [2], [19]. Patients discharged to nursing facilities represent a highrisk cohort requiring enhanced monitoring, structured rehabilitation protocols, and proactive community liaison. The strong predictive importance of chronic comorbidities—hypertension and diabetes—corroborates existing geriatric readmission literature [3], [10] and reinforces the need for integrated chronic disease management pathways extending beyond the inpatient episode.

The precision-recall trade-off observed between XGBoost and logistic regression is a known challenge in imbalanced clinical datasets. XGBoost's superior overall accuracy may give a misleading impression of clinical utility if recall of the minority (readmitted) class remains low. Clinical decision-support applications prioritizing early identification of high-risk patients should favour model configurations that maximize recall, potentially at the expense of precision. Strategies including threshold shifting, SMOTE-based oversampling, or cost-sensitive learning warrant systematic exploration to optimize the recall-accuracy trade-off in this context.

The relatively modest ROC-AUC scores (0.54–0.58) for both models indicate limited discriminative capability. This is consistent with prior studies on imbalanced readmission datasets and suggests that adding longitudinal features, functional status indicators, or socioeconomic variables would likely improve model discrimination. However, the current framework's simplicity confers practical advantages for deployment in resource-constrained Nigerian hospital settings where complex EHR infrastructure is often absent.

V. CONCLUSION

This study presented an XGBoost-based predictive framework for 30-day hospital readmission among geriatric patients, evaluated against a logistic regression baseline. The proposed model achieved higher overall accuracy (68.85%) but lower recall than the baseline, revealing the inherent tension between global classification performance and sensitivity to high-risk minority cases in imbalanced healthcare datasets. Feature importance analysis identified discharge destination, hypertension, and diabetes as the dominant predictors, providing actionable clinical insights that support targeted post-discharge interventions and chronic disease management programs.

The study makes six primary contributions to knowledge: (1) a replicable geriatric readmission prediction framework; (2) practical XGBoost implementation tailored to elderly healthcare data; (3) comprehensive multi-metric model evaluation; (4) empirical prioritization of readmission determinants; (5) demonstration of feasibility in resource-constrained settings; and (6) actionable insights for clinical decision support and care transition planning.

Future work should focus on: (i) integrating real-world longitudinal EHR data from Nigerian hospitals; (ii) incorporating functional status, socioeconomic, and caregiver-related variables; (iii) applying advanced class-imbalance mitigation techniques (SMOTE, cost-sensitive objectives); (iv) validating the framework on external, multi-centre datasets; and (v) deploying the model as a real-time clinical risk stratification tool embedded within hospital information systems.

REFERENCES

[1]. A. Marengoni, S. Angleman, R. Melis, F. Mangialasche, A. Karp, A. Garmen, B. Meinow, and L. Fratiglioni, "Aging with multimorbidity: A systematic review of the literature," Ageing Res. Rev., vol. 10, no. 4, pp. 430–439, 2011.

[2]. K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for healthcare, research, and medical education: A cross-sectional study," The Lancet, vol. 380, no. 9836, pp. 37–43, 2012.

[3]. T. Abiodun, O. Adeyemi, and J. Olawale, "Hospital readmissions among elderly patients: Causes, outcomes, and healthcare implications," African J. Geriatric Med., vol. 12, no. 2, pp. 101–115, 2021.

[4]. R. O. Akinyemi, M. K. Owolabi, A. B. Adeniji, A. O. Afolabi, and R. Ogah, "Stroke in young adults in Nigeria: A single center study," J. Neurol. Sci., vol. 354, no. 1–2, pp. 63–71, 2015.

[5]. A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," N. Engl. J. Med., vol. 380, no. 14, pp. 1347–1358, 2019.

[6]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.

[7]. Y. Wang, B. Kung, and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," Technol. Forecast. Soc. Change, vol. 126, pp. 3–13, 2018.

[8]. S. Chakraborty, A. Roy, and P. Das, "Machine learning approaches for clinical outcome prediction from electronic health records," Healthcare Inform. Res., vol. 25, no. 3, pp. 190–202, 2019.

[9]. R. Kumar, A. Singh, and P. Sharma, "Factors contributing to 30-day readmissions among hypertensive patients: A clinical study," J. Clin. Hypertension, vol. 21, no. 5, pp. 720–729, 2019.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26aug080

[10]. S. Ahmed, A. Musa, and K. Ibrahim, "XGBoost-based prediction of adverse outcomes in older adults using electronic health records," Comput. Biol. Med., vol. 156, Art. no. 106801, 2023.
[11]. I. Okafor, C. Emeka, and O. Abiola, "Predicting hospitalization risk using XGBoost in Nigerian adults," African J. Health Technol., vol. 5, no. 3, pp. 33–48, 2023.
[12]. O. Adegboye, O. Akinola, and J. Ojo, "Machine learning approach to predict hospital readmissions in Nigerian healthcare systems," Nigerian J. Health Inform., vol. 18, no. 2, pp. 45–56, 2022.
[13]. F. Askar, L. Chen, and R. Patel, "Machine learning for all-cause hospital readmission prediction: A systematic review," J. Biomed. Inform., vol. 165, Art. no. 104451, 2024.
[14]. C. González-Nóvoa, L. Pérez, and R. Martínez, "Explainable and optimized machine learning for early ICU readmission prediction," J. Critical Care Inform., vol. 8, no. 2, pp. 45–60, 2023.
[15]. T. Olatunji, O. Adebayo, and H. Musa, "Automated longitudinal feature engineering from EHR for predicting hospital readmission," Comput. Biol. Med., vol. 143, Art. no. 105253, 2022.
[16]. P. Eze, T. Oladipo, and Y. Chen, "Transformer-based integration of clinical notes and structured data for hospital readmission prediction," Artif. Intell. Med., vol. 155, Art. no. 102743, 2025.
[17]. K. W. Johnson, J. Torres Soto, and B. S. Glicksberg, "Artificial intelligence in cardiology: Machine learning applications for hypertension and heart failure management," Nat. Rev. Cardiol., vol. 20, no. 4, pp. 245–261, 2023.
[18]. X. Li, J. Chen, and L. Wang, "Integrating structured and unstructured EHR data for ICU readmission prediction using LSTM and topic modeling," J. Biomed. Inform., vol. 156, Art. no. 104163, 2024.
[19]. H. Yang, Q. Zhang, and K. Liu, "Integrating socioeconomic and clinical features for hospital readmission prediction in hypertension patients," Health Data Sci., vol. 2, no. 3, pp. 99–112, 2022.
[20]. C. Selmer, P. Andersen, and J. Hansen, "Common predictors of 28–31 day readmissions: A review," J. Health Data Sci., vol. 3, no. 1, pp. 20– 38, 2025.