JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Predictive Data Analytics Framework for Hospital Readmission Risk in Chronic Obstructive Pulmonary Disease

[1]Sk Althaf Rahaman, [2]Dr.K.Vedavathi

[1]Research Scholar, [2]Professor
[1]Department of Computer Science, [2]Department of Computer Science
[1]GITAM School of Science, GITAM Deemed to be University, Visakhapatnam, A.P, India
[2]GITAM School of Science, GITAM Deemed to be University, Visakhapatnam, A.P, India

Abstract: Hospital readmission among patients with Chronic Obstructive Pulmonary Disease (COPD) represent a major clinical and economic burden on healthcare systems worldwide. Early identification of patients at high risk of readmission enables proactive interventions, improved care planning, and reduced healthcare costs. In recent years, predictive data analytics and machine learning techniques have been increasingly explored for readmission prediction; however, existing models suffer from limitations such as poor interpretability, inadequate handling of class imbalance, and limited consideration of disease-specific clinical features. This paper presents a comprehensive conceptual framework for COPD patient readmission prediction using predictive data analytics. A systematic review of existing approaches is conducted, followed by the formulation of the readmission prediction problem and identification of relevant clinical and demographic features. Baseline machine learning models, including Logistic Regression and Random Forest, are implemented to establish benchmark performance. Experimental results demonstrate the challenges inherent in COPD readmission prediction and highlight critical research gaps. The findings of this study lay the foundation for the development of novel and enhanced predictive algorithms in future work. Keywords: COPD, Patient Readmission, Predictive Analytics, Machine Learning, Healthcare Data Mining.

Index Terms - **Chronic Obstructive Pulmonary Disease, Hospital Readmission, Predictive Data Analytics, Machine Learning, Electronic Health Records**

I. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a progressive and irreversible respiratory disorder characterized by persistent airflow limitation and recurrent acute exacerbations. It represents a major global health challenge and is consistently ranked among the leading causes of morbidity and mortality worldwide [1]. The chronic nature of COPD, combined with frequent disease exacerbations, results in repeated hospitalizations and long-term healthcare utilization, placing a substantial burden on patients, caregivers, and healthcare systems. Hospital readmissions among COPD patients, particularly those occurring within 30 days of discharge, have emerged as critical indicators of disease severity, quality of care, and effectiveness of post-discharge management strategies. High readmission rates are associated with increased healthcare expenditures, reduced patient quality of life, accelerated disease progression, and elevated mortality risk. Consequently, reducing avoidable hospital readmissions has become a key priority for healthcare providers and policymakers. Accurate prediction of patient readmission risk is essential for enabling early intervention, personalized care planning, and efficient allocation of healthcare resources.

Traditional readmission risk assessment approaches primarily rely on clinician judgment, rule-based scoring systems, or limited statistical models. While clinically useful, these methods often fail to capture the complex and nonlinear interactions among demographic characteristics, clinical variables, comorbidities, medication patterns, and prior hospitalization history that influence readmission risk. The in- creasing adoption of electronic health records (EHRs) has led to the availability of large-scale, high-dimensional healthcare data, creating new opportunities for predictive data analytics. Machine learning techniques offer the ability to model com- plex relationships within such data and have shown promise in predicting hospital readmissions across various chronic dis- eases, including COPD [2]. These data-driven approaches can potentially improve predictive accuracy and support evidence- based clinical decision-making. Despite significant research efforts, existing COPD readmission prediction models exhibit several limitations. Many models demonstrate limited generalizability across patient populations and healthcare settings, while others lack interpretability, restricting their acceptance in clinical practice. Additionally, insufficient incorporation of COPD-specific clinical features and inadequate handling of class imbalance further limit model effectiveness. In this context, the present study aims to address these challenges by presenting a structured conceptual framework for COPD patient readmission prediction using predictive data analytics [3]. A comprehensive review of existing approaches is con- ducted, followed by a baseline predictive analysis to establish benchmark performance.

II. Clinical Background

Chronic Obstructive Pulmonary Disease (COPD) is a major public health issue in India, where it contributes significantly to chronic respiratory morbidity and healthcare utilization. The overall estimated prevalence of COPD in Indian adults is around 7.4 In India, COPD often develops due not only to tobacco smoking but also due to biomass fuel exposure, especially among rural women who cook with wood or dung fuels — a risk factor that disproportionately affects non-smokers and increases disease burden in under-resourced settings. Hospitalization for acute exacerbations of COPD (AECOPD) significantly increases the risk of future readmissions. In an observational study conducted at a tertiary care center in Kerala, nearly 34% COPD patients were readmitted within 30 days of discharge, with prior hospitalizations, a higher BODE score, and poor inhaler technique identified as significant predictors of early readmission2 In the Indian healthcare context, the management of COPD is further complicated by late diagnosis, limited access to pulmonary rehabilitation services, and wide variability in post-discharge follow-up practices across urban and rural settings. Many patients present to tertiary care centers at advanced stages of the disease, often with multiple uncontrolled comorbidities and poor baseline functional status. Additionally, constraints related to healthcare infrastructure, affordability of long-term inhaled therapy, and inconsistent patient education regarding disease self-management contribute to suboptimal continuity of care after hospital discharge. These factors collectively increase vulnerability to early disease exacerbation and un- planned [4]. Consequently, there is a growing clinical need in India for data-driven risk stratification tools that can support clinicians in identifying high-risk COPD patients at the time of discharge and enable targeted interventions, such as closer follow-up, optimized pharmacotherapy, and patient-specific discharge planning.

III. Related Work

Predictive modeling for hospital readmissions in Chronic Obstructive Pulmonary Disease (COPD) has been widely studied in both clinical and data-driven research domains. Clinical observational studies across Asia have identified key risk factors associated with readmission, such as multiple prior admissions, comorbidities, smoking history, eosinophil counts, nutritional deficiencies, and reduced lung function, with 30,90 and 365-day readmission rates reported at approximately 19%, 31%, and 42% respectively across Asian cohorts in systematic analyses [8]. These findings provide valuable guidance for selecting clinically relevant features in predictive models. In the Indian context, several studies have examined COPD readmission determinants and preliminary machine learning models [3]. Prospective observational research conducted in Indian tertiary centres highlighted the influence of comorbid conditions, dyspnea scores, and COPD Assessment Test (CAT) values on early readmissions. Additionally, comparative studies at Indian academic institutions have evaluated the relative performance of classical machine learning classifiers such as Support Vector Machines (SVM), Random Forests (RF), and Decision Trees for COPD readmission prediction, demonstrating feasibility albeit with limited external validation. Smaller scale Indian datasets have also been used to assess demographic and clinical predictors of readmission, though results are constrained by sample size and lack of multicenter representation. Global research has leveraged larger administrative and EHR databases to develop and benchmark predictive models for COPD readmissions [5]. For example, studies using U.S. claims data demonstrated modest improvements in Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) when combining knowledge-driven and data-driven features, though deep learning models did not significantly outperform classical approaches. Other work using nationwide datasets applied machine learning techniques such as Extreme Gradient Boosting (XGBoost), Random Forests, and neural networks, often outperforming standard clinical scoring systems, and identified key predictors including hospitalizations in the previous year and baseline hemoglobin levels. Recent work also highlights that deep learning models such as multilayer perceptron can achieve high sensitivity and specificity in identifying patients at high readmission risk, especially when integrated with rich EHR features [6].

IV. Problem definition

Hospital readmission among patients with Chronic Obstructive Pulmonary Disease (COPD) represents a significant clinical and economic challenge, as early readmissions are often associated with disease severity, inadequate post- discharge care, and suboptimal treatment planning. From a predictive analytics perspective, identifying patients at high risk of readmission enables targeted interventions, optimized resource allocation, and improved patient outcomes. In this study, the COPD readmission prediction problem is formulated as a supervised binary classification task. The objective is to predict whether a discharged COPD patient will experience an unplanned hospital readmission within a predefined time horizon, typically 30 days following discharge. This time window is widely adopted in clinical research and healthcare policy as a benchmark for evaluating quality of care and readmission risk.

Let $X = \{x1, x2. \quad xn\}$ denote the feature vector representing an individual patient, where each feature (xi) corresponds to a demographic attribute (e.g., age, gender), clinical variable (e.g., comorbidities, disease severity scores, laboratory values), or treatment-related factor (e.g., medication usage, length of hospital stay, oxygen therapy). The target variable is defined as $y \{0,1\}$, where (y = 1) indicates hospital readmission within the specified time window and (y = 0) denotes no readmission. Given a dataset $D= \{ (X(i),y(i)) \}$ i=1 comprising m patient records, the goal is to learn a predictive function $f:X \rightarrow y$. that accurately estimates the probability of readmission for new, unseen patients. The predictive model must effectively handle heterogeneous clinical data, potential missing values. and inherent class imbalance due to the relatively lower proportion of readmitted cases. Beyond predictive accuracy, an important consideration in this problem is clinical interpretability. Models should provide meaningful insights into the factors contributing to readmission risk, enabling clinicians to understand, trust, and potentially act upon model predictions. Therefore, the primary objective of this study is to develop and evaluate baseline predictive models that achieve a balance between classification performance, robustness, and interpretability, thereby establishing a foundation for subsequent algorithmic enhancement and optimization in future research.

V. Dataset Description

This study utilizes a publicly available critical care dataset derived from electronic health records (EHRs), such as the Medical Information Mart for Intensive Care IV (MIMIC- IV). The dataset contains de-identified longitudinal health records of patients admitted to intensive care units, including demographic information, diagnoses, laboratory test results, vital signs, medications, procedures, and hospitalization details [7].

TABLE I
Feature Categories Used in the Study

Feature Category	Example Variables
Demographic	Age, Gender
Clinical	COPD severity indicators, Comorbidities
Laboratory	Hemoglobin, White Blood Cell Count
Vital Signs	Heart Rate, Respiratory Rate HospitalizationLength of Stay, Prior Admissions
Outcome	30-day Readmission Status

Patients diagnosed with Chronic Obstructive Pulmonary Disease (COPD) were identified using standardized ICD-9 and ICD-10 diagnosis codes recorded during hospital admissions. Only adult patients aged 18 years and above were included in the analysis. For patients with multiple admissions, the first eligible hospitalization was considered as the index admission to prevent data leakage across samples.

TABLE II
Patient-Level Dataset

ID	Age	Gen	LOS	Prev	Hb	Readm
P001	68	M	7	2	11.8	1
P002	72	F	5	1	12.5	0
P003	65	M	9	3	10.9	1
P004	59	M	4	0	13.2	0
P005	74	F	8	2	11.1	1

where LOS denotes length of hospital stay (days), Prev indicates the number of prior admissions, Hb represents hemoglobin concentration (g/dL), and Readm denotes 30-day readmission status. Hospital readmission was defined as any unplanned inpatient admission occurring within 30 days following discharge from the index hospitalization. Planned admissions and inter-hospital transfers were excluded. Each patient record was labeled with a binary outcome variable indicating readmission (1) or no readmission (0).

TABLE III
Binary Outcome Definition for Readmission Prediction

Outcome Value	Description
1	Patient readmitted within 30 days of discharge
0	Patient not readmitted within 30 days of discharge

Patients with COPD were identified using standardized diagnosis codes from the International Classification of Dis- eases (ICD), as recorded in the dataset. Patients with COPD were identified using standardized diagnosis codes from the International Classification of Diseases (ICD), as recorded in the dataset.

1) ICD-9 Codes:

- 490 – Bronchitis, not specified as acute or chronic
- 491.0 – Simple chronic bronchitis
- 491.1 – Mucopurulent chronic bronchitis
    - 491.20 – Obstructive chronic bronchitis without exacerbation
    - 491.21 – Obstructive chronic bronchitis with acute exacerbation
    - 492.0 – Emphysematous bleb
    - 492.8 – Other emphysema
    - 496 – Chronic airway obstruction, not elsewhere classified

2) ICD-10 Codes:

- J40 – Bronchitis, not specified as acute or chronic
- J41.0 – Simple chronic bronchitis
- J41.1 – Mucopurulent chronic bronchitis
- J42 – Unspecified chronic bronchitis
- J43.0 – MacLeod's syndrome
- J43.1 – Panlobular emphysema
- J43.2 – Centrilobular emphysema
- J43.8 – Other emphysema
- J43.9 – Emphysema, unspecified
- J44.0 – COPD with acute lower respiratory infection
- J44.1 – COPD with acute exacerbation, unspecified
- J44.9 – Chronic obstructive pulmonary disease, unspecified

These ICD-9 and ICD-10 codes are widely adopted in clinical and epidemiological research and enable consistent, reproducible identification of COPD patient cohorts across healthcare datasets. These codes are widely adopted in clinical research and ensure consistent and reproducible identification of COPD patients across studies.

VI. Methodology

This study adopts a structured predictive analytics method- ology to establish baseline performance for COPD readmission prediction. The workflow includes data preprocessing, feature representation, baseline model development, and performance evaluation, with emphasis on reproducibility and clinical interpretability [3].

A. *Data Preprocessing*

Electronic health record data contain missing values, heterogeneous feature scales, and noise. Continuous variables were imputed using mean or median values based on their distributions, while categorical variables were encoded using label or one-hot encoding. Feature scaling was applied to continuous attributes to reduce scale bias. Records with in- complete outcome labels or implausible values were excluded to ensure data integrity.

B. *Feature Selection and Representation*

The feature set includes demographic, clinical, laboratory, and hospitalization-related variables selected based on clinical relevance and data availability. Key features include age, gender, length of stay, prior admissions, selected laboratory measurements, and comorbidity indicators. To maintain interpretability, no aggressive dimensionality reduction techniques were applied in this baseline study.

C. *Baseline Predictive Models*

Three supervised learning models were implemented to benchmark readmission prediction performance.
Logistic Regression (LR) was used as an interpretable statistical baseline model. *Support Vector Machine (SVM)* was employed to capture potential nonlinear relationships among features. *Random Forest (RF)* was utilized as an ensemble model capable of handling heterogeneous clinical data and providing feature importance estimates.

D. *Handling Class Imbalance*

To address class imbalance, stratified train–test splitting was applied to preserve class proportions. Class-weighting strategies were incorporated into applicable models to improve sensitivity toward the minority readmission class [8].

E. *Model Training and Evaluation*

Models were trained on the training subset with hyperparameters selected using default configurations or grid search. Cross-validation was applied within the training set, and final performance was evaluated on a held-out test set using accu- racy, precision, recall, F1-score, and AUC-ROC. Emphasis was placed on recall and AUC-ROC due to their clinical relevance

VII. Results and Discussion

This section presents the experimental results obtained from the baseline predictive models and discusses their implications for COPD readmission prediction. Model performance was evaluated on a held-out test set using standard classification metrics.

A. *Performance Comparison of Baseline Models*

Logistic Regression achieved moderate predictive performance while offering strong interpretability. The Support Vector Machine demonstrated improved discrimination capability by modeling nonlinear relationships among clinical features. Random Forest achieved the best overall performance across most evaluation metrics, highlighting the effectiveness of ensemble-based learning for heterogeneous electronic health record data.

Model	Acc	Prec	Rec	F1	AUC
Logistic Regression	0.71	0.64	0.58	0.61	0.73
Support Vector Machine	0.74	0.67	0.61	0.64	0.76
Random Forest	0.78	0.71	0.66	0.68	0.81

B. *Discussion*

The superior performance of Random Forest can be at- tributed to its ability to capture complex interactions between clinical and hospitalization-related features without strong assumptions regarding data distribution. In contrast, Logistic Regression provides transparent risk estimation but is limited in modeling nonlinear patterns. While the Support Vector Ma- chine improves predictive accuracy, its reduced interpretability may limit direct clinical adoption. Despite performance improvements, all models exhibit limitations in sensitivity toward the minority readmission class, reflecting the inherent class imbalance in hospital readmission data. These findings emphasize the need for advanced imbalance-aware learning strategies and clinically informed feature engineering.

Overall, the results establish benchmark performance levels and validate the feasibility of predictive analytics for COPD readmission risk assessment. The observed limitations motivate future research focused on algorithmic innovation, enhanced interpretability, and improved generalization across diverse patient populations.

VIII. Conclusion and Future Work

This paper presented a structured baseline framework for predicting 30-day hospital readmission among patients with Chronic Obstructive Pulmonary Disease (COPD) using predictive data analytics. By formulating the readmission task as a binary classification problem and leveraging clinically relevant features derived from electronic health records, this study established benchmark performance levels using widely adopted machine learning models. Experimental results demonstrate that ensemble-based methods, particularly Random Forest, outperform traditional statistical models in capturing complex relationships among demographic, clinical, and hospitalization-related variables. However, the observed performance remains constrained by class imbalance, limited temporal representation, and challenges related to clinical interpretability. These findings high- light the need for more specialized modeling approaches tailored to COPD-specific disease characteristics. Future work will focus on the development of novel COPD- specific predictive algorithms that incorporate advanced feature engineering, temporal modeling, and imbalance-aware learning strategies [9]. Additional efforts will be directed toward improving model explainability and validating performance across diverse patient cohorts. These extensions will form the basis for subsequent research contributions and the final doctoral thesis, with the ultimate goal of supporting clinically actionable decision-making for COPD readmission risk management.

REFERENCES

[1] World Health Organization, “Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach,” World Health Organization, 2007.
[2] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.
[3] Sk. Althaf Rahaman, “Data analytics foundations, challenges and data exploration process,” High Technology Letters, vol. 27, no. 6, pp. 21–25, 2021.
[4] Global Initiative for Chronic Obstructive Lung Disease, “Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease,” GOLD Report, 2023.
[5] S. J. Shah and M. C. Fang, “Predicting hospital readmissions using electronic health records,” Journal of General Internal Medicine, vol. 31, no. 2, pp. 1–8, 2016.
[6] A. Esteva, A. Robicquet, and B. Ramsundar, “A guide to deep learning in healthcare,” Nature Medicine, vol. 25, pp. 24–29, 2019.
[7] A. E. W. Johnson et al., “Mimic-iv, a freely accessible electronic health record dataset,” Scientific Data, vol. 8, no. 1, pp. 1–9, 2021.
[8] N. V. Chawla, K. W. Bowyer, and L. O. Hall, “Smote: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[9] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, vol. 30, 2017.
[10] S. Alzahrani et al., “Predictive modeling of hospital readmissions for patients with chronic obstructive pulmonary disease,” International Journal of Medical Informatics, vol. 118, pp. 64–72, 2018.
[11] Y. Wang and L. Wang, “A machine learning approach to predict 30-day hospital readmissions among copd patients,” BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1–12, 2020.
[12] R. Miotto, F. Wang, and S. Wang, “Deep learning for healthcare: review, opportunities and challenges,” Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.

[1] [4] [3] [5] [2] [6] [8] [7] [9] [10] [11] [12]