Saudi Toxicology Journal (STJ)

Journal home page: https://uqu.edu.sa/s.toxicology.s/S.T.J

Article

Machine Learning–Based Prediction of Potentially Inappropriate Prescribing in Geriatric Patients Presenting to the Emergency Department at Al-Qatif Central Hospital, Saudi Arabia

Aymen Alqurain[1]*

[1]Department of Clinical Practice, Faculty of Pharmacy, Northern Border University, Rafha 91911, Saudi Arabia;
* Correspondence: aymen.alqurain@nbu.edu.sa

Citation: Alqurain, A Machine Learning–Based Prediction of Potentially Inappropriate Prescribing in Geriatric Patients Presenting to the Emergency Department at Al-Qatif Central Hospital, Saudi Arabia. *STJ*, **2025**, 1 (3),48-64. https://doi.org/10.70957/uqu.edu.sa/s.toxicology.s/stj.2025.1.3.4

Received: 19 Nov 2025

Accepted: 19 Dec 2025

Published: 23 Dec 2025

Abstract

Potentially inappropriate prescribing (PIP) is an avoidable factor leading to adverse drug events in older adults, particularly in the emergency department (ED). This study aimed to train and validate a machine learning (ML) model to predict the probability of PIP in elderly patients in the ED. This retrospective analysis encompassed patients aged 65 years and above who presented to the Al-Qatif Central Hospital. PIPs were identified using a validated screening tool designed for older persons to ensure appropriate treatment. Variables have been detected by least absolute shrinkage and selection operator regression. An extreme gradient boosting model was trained and assessed across training, internal validation, and external validation cohorts. The evaluation of model performance was performed using receiver operating characteristic areas under the curve (ROC-AUC) and decision curve analysis. Among 4942 patients, 80% exhibited at least one PIP. The final model exhibited robust discrimination in the external validation cohort (ROC-AUC = 0.7, accuracy = 0.84, Brier score = 0.13). Important factors contributing to PIP risk were cardiovascular disease, the number of medications prescribed in the ED, arthritis-related diseases, and musculoskeletal pain. The model exhibited effective calibration and offered clinical value across a broad spectrum of thresholds (20 – 60%). This model demonstrated robust predictive efficacy in detecting PIP risk among older ED patients. Its implementation may augment prescribing safety and promote clinical outcomes in geriatric care.

Keywords: Potentially Inappropriate Prescribing, Older Patients, Emergency Departments, Machine Learning, STOPP/START Criteria, Predictive Modeling

1. Introduction

Geriatric patients face increased risk of medication-related problems due to comorbidities, age-associated alterations in pharmacokinetics and pharmacodynamics, and the prevalent use of multiple medications, commonly referred to as polypharmacy [1]. This population is particularly vulnerable to adverse drug events, functional deterioration, and hospital readmissions due to inappropriate prescribing practices [2]. A primary factor in these concerns is the use of potentially inappropriate medications (PIMs), which are drugs that pose a higher risk than benefit when safer or more effective alternatives exist. Various tools, such as the Screening Tool of Older Persons' Prescriptions (STOPP) and the Screening Tool to Alert to Right Treatment (START) and Beer's Criteria, have been established to detect PIMs and improve prescribing safety in older populations [3, 4].

However, in fast-paced clinical settings like the emergency department (ED), the successful utilization of prescribing tools is often challenging [5]. Older patients often arrive at the ED with complex diseases and insufficient or inaccessible medical histories. Clinicians often must make immediate treatment decisions without extensive information or detailed medication reviews. These issues increase the risk of prescribing errors and elevate the probability of potentially inappropriate prescribing (PIP) during emergency care [6]. Recent studies indicate a significant incidence of PIP, with almost 90% of community-dwelling older patients and 76% of older ED patients receiving at least one PIM [7, 8].

Although tools like STOPP/START and the Beers Criteria provide valuable guidance, their reliance on manual assessment and clinical judgment constrains their regular implementation in the ED. Therefore, an automated, quick solution is required for the integration into standard electronic health records (EHRs) [9, 10]. In this context, machine learning (ML) techniques have arisen as powerful approaches to improve medication safety [11]. These models can analyze extensive patient data, identify complex patterns, and predict risks in real-time, thereby assisting clinicians in making more secure prescribing choices.

Previous efforts to develop and train predictive models for medication safety have predominantly concentrated on inpatient or long-term care settings [10, 12, 13]. These models generally depend on conventional statistical methods and focus exclusively on the existence or absence of PIP. Most of these models lack external validation and disregard the severity or quantity of PIMs. Importantly, limited research has focused on the ED setting, despite its high-risk nature and the distinctive challenges it poses for prescribing decisions. Accordingly, this study is designed as a context-specific validation and implementation study, aimed to train and validate a ML model to predict the probability of PIP among older adults attending the ED. By focusing on the ED, the study addresses a significant gap in medication safety research, as the final model will offer a practical and scalable approach to facilitate safer prescribing decisions for older adults in the ED.

2. Materials and Methods

2.1. Study Design and Characteristics of Study Population

This retrospective cross-sectional study carried out at Al-Qatif Central Hospital (QCH), a leading referral facility in the Eastern Province of Saudi Arabia, from January 2020 to December 2021. Eligible participants included patients aged 65 years or older who visited the ED during the study period. Patients were excluded if they were younger than 65 years, visited outpatient clinics, or were admitted to inpatient medical, surgical, or intensive care units. The study employed data solely from the first documented visit of individuals with multiple ED visits.

A subset of patients who visited the ED during the final three months of the study period (n = 449) was designated for external validation. **Figure 1** presents a summary of the patient enrollment procedure and model training and validation methodology.

2.2. Data Collection and Variables Definitions

Demographic data, morbidities, laboratory tests findings, and the prescribed medications list were collected from electronic health and pharmacy records.

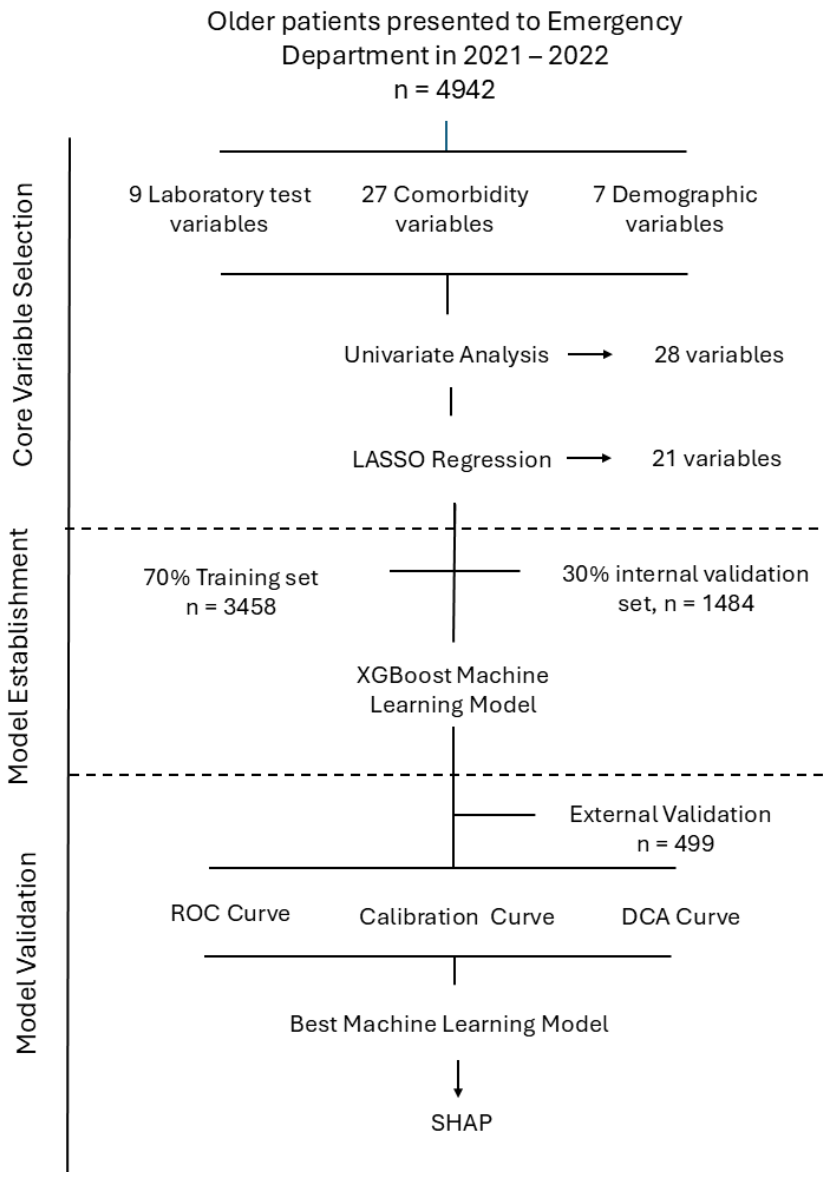

Figure 1. Flowchart for patient enrolment and model training and validation steps. *LASSO: stands for Least Absolute Shrinkage and Selection Operator regression, XGBoost stands for Extreme Gradient Boosting, ROC-AUC stands for Receiver Operating Characteristic Area Under the Curve, DCA stands for Clinical Decision Curve, and SHAP stands for SHapley Additive exPlanations*

Comorbidities were determined through medical records documentation or by utilizing the Rx-Risk comorbidity index on the long-term medication list and thereafter classified according to the International Classification of Diseases, 10th Version (ICD-10) [14, 15]. The Charlson Comorbidity Index (CCI) was employed to evaluate the 1-year mortality risk, whereas creatinine clearance (CrCl) was calculated using the Cockcroft-Gault equation [16, 17].

To reduce diagnostic overlap and decrease the probability of model overfitting, clinically relevant disorders were classified as follows: asthma and chronic obstructive pulmonary disease (COPD) categorized as "respiratory diseases"; osteoarthritis and rheumatoid arthritis classified under "arthritis-related diseases"; divers forms of pain in the lower and upper back, muscles, bones, joints, ligaments, and tendons identified as "musculoskeletal pain"; peptic ulcer, gastroesophageal reflux disease, and other forms of

gastritis disorder recognized as "gastritis-related diseases"; and hypertension, ischemic heart diseases, heart failure, and arrhythmia designated as "cardiovascular diseases".

Medication data comprised prescriptions issued during ED visits as well as patients' regularly prescribed medications, encompassing long-term, short-term, as-needed prescriptions, and supplements. All medications were categorized according to the Anatomical Therapeutic Chemical (ATC) system [18].

2.3. Identification and Classification of Potential Inappropriate Prescribing

PIPs were identified using the STOPP/START criteria, version 3 [3]. This tool is widely employed for patients aged 65 years and above in both acute and outpatient facilities. The overall number of medications prescribed at the ED and the total number of regular long-term prescriptions were recorded and reported. The principal investigator manually applied the STOPP/START criteria to identify PIMs. To minimize bias and ensure classification precision, a random sample of patient data was evaluated by two independent reviewers to validate PIM identification.

Each STOPP or START violation was recorded as a single incident, with over-prescribing and under-prescribing given equal weight in the assessment of the overall PIP burden. This inclusive strategy was implemented to capture the intricacies of prescribing decisions in the ED, where both improper medication utilization and the omission of necessary therapy elevate patient risk. Additionally, the total number of PIPs identified per patient was calculated to quantify the burden of PIP prescribing. This measure indicates cumulative exposure to PIP rather than direct clinical harm. An elevated PIM count indicated an increased level of prescribing risk, allowing the model to encompass both a binary classification and a spectrum of inappropriate prescribing risk.

2.4. Preliminary Statistical Analysis and Significant Variable Identification

Descriptive statistics were reported as mean ± standard deviation (SD) for normally distributed data, median (interquartile range (IQR)) for skewed data, and frequencies (%) for categorical data. Group comparisons were conducted utilizing the Student t-test, Mann-Whitney U test, or chi-square test, as appropriate. Variables with a p-value less than 0.05 were considered statistically significant and selected for inclusion in the model training process.

The selection of predictors involved a two-step process: variables exhibiting significant differences in univariate analysis ($p < 0.05$) were incorporated into a Least Absolute Shrinkage and Selection Operator (LASSO) regression utilizing 10-fold cross-validation. This approach successfully reduced overfitting and determined the most predictive feature set [9].

2.5. Model Training and Performance Evaluation

The Extreme Gradient Boosting (XGBoost) model was chosen due to its versatility and widespread application in contemporary clinical prediction, as reported in previous studies [19, 20]. The XGBoost model is an ensemble learning technique that produces classifications by pooling predictions from many decision trees. Its advantages encompass resilience in handling high-dimensional data, outstanding prediction precision, and resistance to overfitting. The XGBoost model demonstrates exceptional efficacy with extensive datasets, delivering enhanced accuracy and reliability [21].

The evaluation of model performance was conducted throughout training, internal, and external validation sets, emphasizing discrimination, calibration, and clinical utility. This study employed two validation methods that were based on the final predictor used. The first predictor was initially established as proposed earlier, measuring risk as a continuous outcome. This approach utilized root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2) to evaluate the model performance. In the second approach, the predictor was classified as a binary result (present/absent). This method evaluated discriminative capability by the Receiver Operating Characteristic Area Under the Curve (AUC-ROC). The assessment of model calibration was performed using a calibration curve created through 1,000 bootstrap resampling

iteration which evaluates the alignment between projected probability and observed outcomes, depicted as scatter plots comparing expected and actual event rates. Participants were arranged in ascending order according to predicted probability and subsequently divided into equal groups. The mean predicted probability and the observed event rate were calculated and plotted for each group, with the expected probability represented on the x-axis and the observed event rate on the y-axis. The Brier score, which quantifies the mean squared error between predicted probabilities and actual outcomes, was utilized to assess model performance; a lower score signifies enhanced model efficacy [9].

2.7. Assessment of Clinical Utility and Net Benefit of the Final Model

Clinical decision curve analysis (DCA) was utilized to evaluate clinical net benefit and confirm the model's clinical relevance. This analysis assesses the utility and cost-effectiveness of the prediction model by determining the optimal threshold probability, calculating net benefits, and guiding clinical decision-making [9].

In DCA, two distinct scenarios are assessed: treating all patients or treating none. A prediction model is deemed medically beneficial only if its curve surpasses both scenarios. The optimal threshold probability can be found by net benefit analysis, thereby enhancing informed clinical decision-making and directing suitable therapies.

2.8. Visualization of the Final Model

The SHapley Additive exPlanations (SHAP) method was employed to create a bee-swarm plot, illustrating the contribution of each characteristic to the predicted outcomes. Furthermore, SHAP force charts were generated for specific examples to demonstrate the impact of distinct attributes on predictions for each sample, so offering a more profound understanding of the model's decision-making process.

2.9. Ethical Approval

The data utilized for this study was gathered following the approval from the Institutional Review Board (IRB) at Mohammed Al-Mana College for Medical Sciences (SR/RP/79, Dated 17/02/2022) and the IRB at QCH (QCH-SREC019/2022, Dated 08/06/2022).

3. Results

3.1. Characteristics of the Enrolled Patients

A total of 4,942 patients aged 65 year and older were incorporated in the preliminary model training phases, comprising 2,348 (48%) males and 2,594 (52%) females (**Table 1**). The median age was 73 years (IQR: 68 – 80). The primary disease categories comprised arthritis-related diseases (71%), gastritis-related disorders (47%), coagulation disorders (43%), and musculoskeletal pain (43%) (**Table 1**). A total of 3960 patients (80%) were prescribed at least one PIM, as determined by the STOPP/START criteria, version 3. **Table 1** illustrates that patients in the PIM group exhibited significantly elevated CCI ratings (5 vs. 3, $p < 0.001$), received a greater number of medications in the ED (3 vs. 2, $p < 0.001$), and had more frequent long-term prescriptions in their profiles (4 vs. 3, $p < 0.001$) compared to the non-PIM group. The PIM group demonstrated a higher prevalence of multiple comorbidities, including epilepsy (9% vs. 2%, $p < 0.001$), depression (4% vs. 1%, $p < 0.001$), diabetes mellitus (29% vs. 5%, $p < 0.001$), constipation (20% vs. 13%, $p < 0.001$), and anemia (29% vs. 11%, $p < 0.001$) relative to the non-PIM group (**Table 1)**.

Table 1. Patients' characteristics classified as gender differences.

Characteristics	Entire Cohort	PIM Group	Non-PIM Group	p - value
	n = 4942	n = 3960	n = 982	
Age, median (IQR)	73 (68 – 80)	73 (68 – 80)	72 (68 – 79)	0.024
Weight, mean (SD)	73 (17)	73 (17)	72 (17)	0.08
CrCl, mean (SD)	73 (31)	73 (31)	74 (30)	0.7
CCI, mean (SD)	5 (3 – 7)	5 (4 – 7)	3 (3 – 4)	< 0.001
NPM at emergency, median (IQR)	3 (2 – 5)	3 (2 – 5)	2 (1 – 3)	< 0.001
NPM as regular, median (IQR)	4 (2 – 8)	4 (2 – 9)	3 (2 – 7)	< 0.001
Sex (Male), n (%)	2348 (48)	1852 (47)	496 (51)	0.04
(Female), n (%)	2594 (52)	2108 (53)	486 (49)	
Anxiety, n (%)	322 (7)	271 (7)	51 (5)	0.06
Dementia, n (%)	24 (1)	22 (1)	2 (0.2)	0.2
Epilepsy, n (%)	393 (8)	370 (9)	23 (2)	< 0.001
Depression, n (%)	174 (4)	163 (4)	11 (1)	< 0.001
Parkinson, n (%)	56 (1)	50 (1)	6 (1)	0.8
Psychiatric diseases, n (%)	128 (3)	115 (3)	13 (1)	0.005
Diabetes mellitus, n (%)	1204 (24)	1153 (29)	51 (5)	< 0.001
Thyroid disorders, n (%)	224 (5)	210 (5)	14 (1)	< 0.001
Gastritis related disease, n (%)	2342 (47)	2141 (54)	201 (21)	< 0.001
Constipation, n (%)	927 (19)	796 (20)	131 (13)	< 0.001
Liver diseases, n (%)	609 (12)	523 (13)	86 (9)	< 0.001
Emesis disorder, n (%)	434 (9)	364 (9)	70 (7)	0.04
Coagulation disorder, n (%)	2099 (43)	1998 (51)	101 (10)	< 0.001
Renal diseases, n (%)	173 (4)	167 (4)	6 (1)	< 0.001
Anemia, n (%)	1235 (25)	1131 (29)	104 (11)	< 0.001
Arrhythmia, n (%)	160 (3)	159 (4)	1 (0.1)	< 0.001
Heart failure, n (%)	1741 (35)	1737 (44)	4 (0.4)	< 0.001
Hypertension, n (%)	1810 (37)	1801 (46)	9 (1)	< 0.001
Ischemic heart disease, n (%)	1974 (40)	1974 (50)	2 (0.2)	< 0.001
Lipidemia, n (%)	1551 (31)	1504 (38)	47 (5)	< 0.001
Incontinence, n (%)	590 (12)	522 (13)	68 (7)	< 0.001
Infections, n (%)	2021 (41)	1760 (44)	261 (27)	< 0.001
Gout, n (%)	93 (2)	91 (2)	2 (0.2)	< 0.001
Muscular Pain, n (%)	2126 (43)	1987 (50)	139 (14)	< 0.001
Respiratory disorder, n (%)	657 (13)	608 (15)	49 (5)	< 0.001
Arthritis related disease, n (%)	3500 (71)	2787 (70)	713 (73)	0.2
Osteoporosis, n (%)	1164 (24)	1083 (27)	81 (8)	< 0.001
Vitamin D3, mean (SD)	119 (55)	119 (55)	121 (54)	0.3
Albumin, mean (SD)	40 (9)	40 (9)	40 (9)	0.5
Alanine Transferase, mean (SD)	32 (15)	32 (15)	32 (17)	0.9
Aspartate Transferase, mean (SD)	21 (16)	20 (14)	21 (21)	0.5
Cholesterol, mean (SD)	3.6 (1.3)	3.6 (1.3)	3.5 (1.3)	0.07
Hemoglobin, mean (SD)	13.6 (2.4)	13.6 (2.4)	13.6 (2.7)	0.7
Random Glucose, mean (SD)	7.1 (3.3)	7.1 (3.3)	7 (3.2)	0.5
International normalized ratio, mean (SD)	1.2 (0.5)	1.2 (0.5)	1.2 (0.6)	0.5
Potassium, mean (SD)	4.4 (0.6)	4.4 (0.6)	4.4 (0.6)	0.08

SD: Standard Deviation, IQR: Interquartile Range, CrCl: Creatinine Clearance, CCI: Charlson Comorbidity Index

3.2. Pattern of Potentially Inappropriate Medication Prescription

The mean number of PIMs per patient was 3.4 ± 3.0. Female patients demonstrated a significantly greater mean number of PIMs than male patients (3.5 vs. 3.2, p = 0.001) (**Table 2**). Female patients received prescriptions for medications indicated by the START criteria more often than male patients (1.9 vs. 1.7,

Table 2. Prevalence and pattern of potentially inappropriate medications within the study cohort, categorized by gender.

Characteristics	Entire Cohort	Male group	Female group	p - value
	n = 4942	n = 2348	n = 2594	
PIM, mean (SD)	3.4 (3)	3.2 (3)	3.5 (3)	0.001
STOPP, mean (SD)	1.5 (1.8)	1.5 (1.8)	1.6 (1.9)	0.07
START, mean (%)	1.8 (1.8)	1.7 (1.8)	1.9 (1.9)	< 0.001
PIM CVD, mean (SD)	0.6 (0.9)	0.6 (0.9)	0.6 (1)	0.5
PIM Coagulation	0.13 (0.5)	0.15 (0.5)	0.11 (0.4)	0.001
PIM Gastroenterology	0.12 (0.3)	0.1 (0.3)	0.13 (0.4)	0.001
PIM Respiratory	0.01 (0.1)	0.01 (0.1)	0.01 (0.1)	0.8
PIM Musculoskeletal	0.35 (0.6)	0.33 (0.6)	0.37 (0.6)	0.001
PIM Endocrinology	0 (0.04)	0 (0.03)	0 (0.05)	0.5
PIM Fall	0.24 (0.5)	0.22 (0.5)	0.26 (0.6)	0.007
PIM Analgesic	0.09 (0.4)	0.08 (0.4)	0.9 (0.4)	0.007

PIM: Potentially Inappropriate Medication, SD: Standard Deviation, STOPP: Screening Tool of Older Persons' Prescriptions, START: Screening Tool to Alert to Right Treatment, CVD: Cardiovascular Diseases.

Table 3. Commonly recorded potentially inappropriate prescribed medications in the study population, classified by gender.

criterion	Entire Cohort	Male Group	Female Group	p - value
	n = 4942	n = 2348	n = 2594	
BBs as mono therapy for HTN	115 (2)	53 (2)	62 (2)	0.8
QT prolongation medications	1048 (19)	530 (21)	518 (18)	0.01
NSAID with IHD	1275 (23)	589 (23)	686 (24)	0.5
NSAID with HF	500 (9)	197 (8)	303 (11)	< 0.001
ASA with clopidogrel	156 (3)	89 (4)	67 (2)	0.01
Antiplatelet with anticoagulant	277 (5)	162 (6)	115 (4)	< 0.001
Asa with no CVD	131 (2)	71 (3)	60 (2)	0.1
PPI long therapy	426 (8)	173 (7)	253 (9)	0.006
Corticosteroid with GORD	130 (2)	54 (2)	76 (3)	0.2
NSAID with no PPI	1302 (24)	555 (22)	747 (26)	< 0.001
NSAID for OA with no Paracetamol	196 (4)	96 (4)	100 (4)	0.6
Opioid for OA	502 (9)	228 (9)	274 (10)	0.5
Opioid with no laxative	654 (12)	274 (11)	380 (13)	0.005

BBs: Beta Blockers, HTN: Hypertension, NSAIDs: Non-Steroidal Anti-Inflammatory Drugs, IHD: Ischemic Heart Diseases, HF: Heart Failure, ASA: Aspirin, CVD: Cardiovascular diseases, PPI: Proton Pump Inhibitors, GORD: Gastro-Oesophageal Reflux Diseases, OA: Osteoarthritis.

$p < 0.001$). Marked gender disparities were observed in gastrointestinal and musculoskeletal-related PIM, with female patients demonstrating a greater average usage than their male counterparts (**Table 2**).

The most commonly identified PIMs included non-steroidal anti-inflammatory drugs (NSAIDs) without proton pump inhibitors (PPI) protection (24%), QT-prolonging medications (mostly macrolide antibiotics and amiodarone) (19%), NSAID use in patients with ischemic heart disease (23%), opioids without concurrent laxative use (12%), and dual antiplatelet therapy (aspirin + clopidogrel) (3%) (**Table 3**).

3.3. Significant Clinical Variable Idenitfication

Out of 28 variables significantly associated with PIM prescribing in the univariate analysis, 21 were retained by LASSO regression with a penalty parameter (lambda (λ) = 0.0011). The principal predictors identified encompassed age, gender, number of medications prescribed in the ED, and the existence of comorbidity, including epilepsy, depression, psychotic

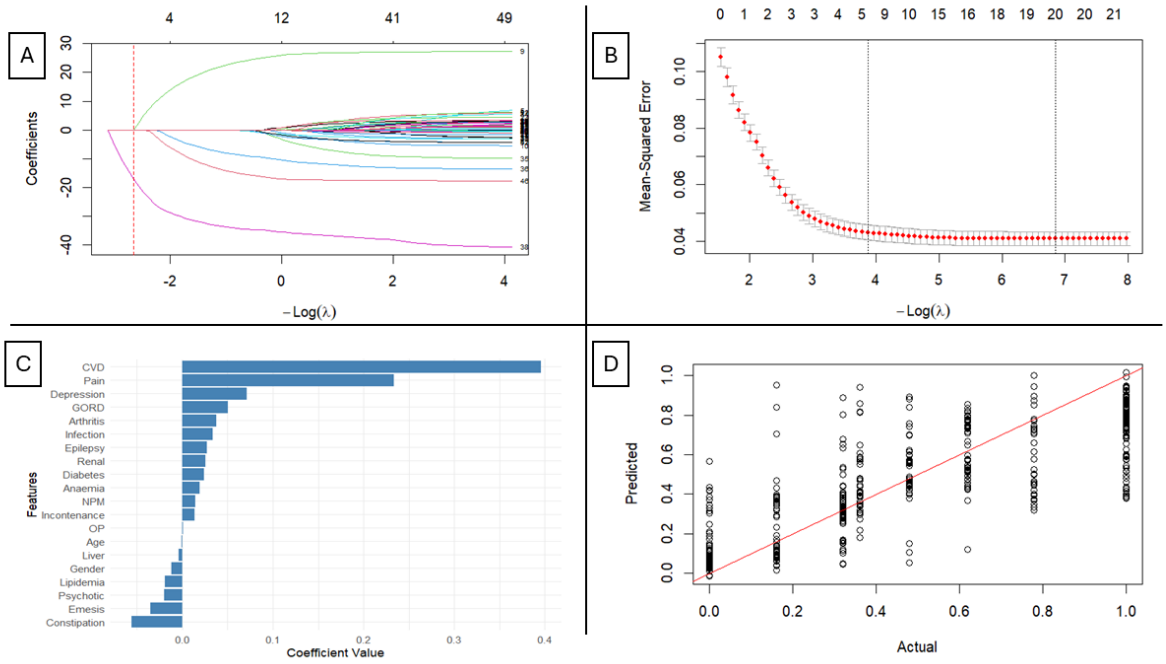

Figure 2. Feature selection and model calibration diagnostics plots. *Panel A represents LASSO coefficient path plot; Panel B represents cross-validation curve (mean Squared Error vs. λ); and Panel C represents LASSO-selected features and their coefficients, and Panel D: predicted versus actual plot (model fit). Λ: lambda.*

Table 4. Model evaluation results for potentially inappropriate prescribing (PIP) burden risk prediction.

Parameter	Training	Internal	External
RMSE	0.16	0.18	0.22
MAE	0.12	0.14	0.16
R2	0.77	0.66	0.62

RMSE: Root mean square error, MAE: Mean absolute error, R^2: R-squared.

disorders, diabetes mellitus, gastritis-related disorders, liver diseases, constipation, emesis, renal diseases, anemia, cardiovascular diseases, dyslipidemia, urinary incontinence, infection, musculoskeletal pain, respiratory disorders, arthritis-related diseases, and osteoporosis (**Figure 2**).

3.4. Training and Validation of the Final Model

A total of 4,942 patients were randomly assigned to a training set (n = 3,458; 70%) and an internal validation set (n = 1,484; 30%). An XGBoost ML model was trained to predict PIP risk using the clinical variables determined by LASSO regression. The efficacy of the XGBoost model in predicting PIP burden risk was assessed utilizing three principal metrics (RMSE, MAE, and R2) throughout the training, internal, and external validation sets, with the findings summarized in **Table 4**.

The model was further evaluated to yield a binary predictor (yes/no) for the existence of PIP. The model demonstrated a ROC-AUC of 0.8 (SD; 0.02) in the training set, 0.076 (SD; 0.01) in the internal validation, and 0.7 (SD; 0.04) in the external validation (**Figure 3**). The confusion matrix plots for the model demonstrate a balanced predictive performance in the training set, with 153 false positives and 149 false negatives (**Figure 4**). A comparable pattern was noted in the internal validation set, comprising 84 false positives and 88 false negatives. In the external validation set, the number of false negatives exceeded that of false positives (48 compared to 17). In total, these misclassified instances constituted 13% of the external validation set, while 87% of the predictions were accurate.

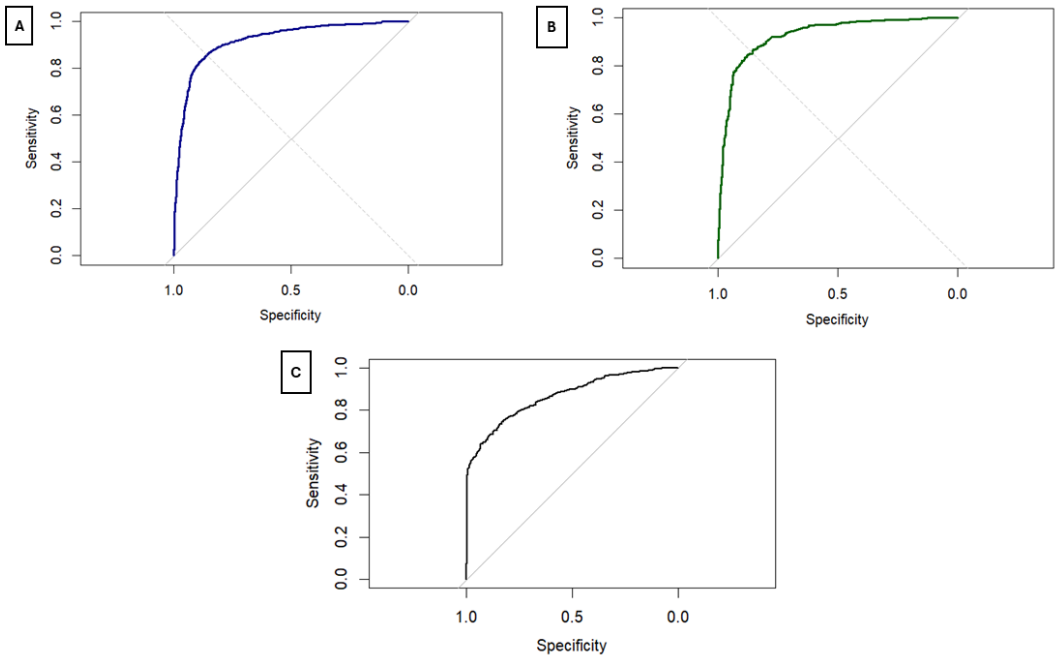

Figure 3. The ROC-AUC for the final trained model in (A) training, (B) internal validation, and (C) external validation datasets.

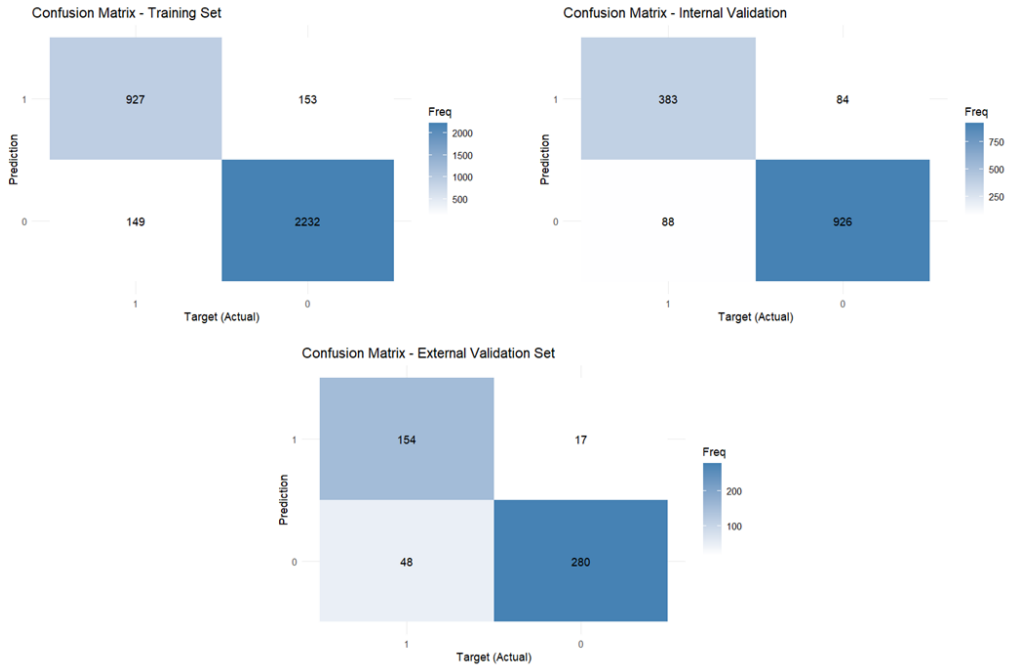

Figure 4. Confusion matrices illustrating the performance of the final model throughout training, internal validation, and external validation datasets.

The calibration curve demonstrated a strong correlation between the expected and actual probabilities of PIP (**Figure 5**). In the external validation set, the model attained a Brier score of 0.13, indicating proficient model calibration. DCA plots further demonstrated that the model offered a greater net therapeutic benefit compared to the "treat-all" and "treat-none" strategies within a clinically pertinent probability threshold range

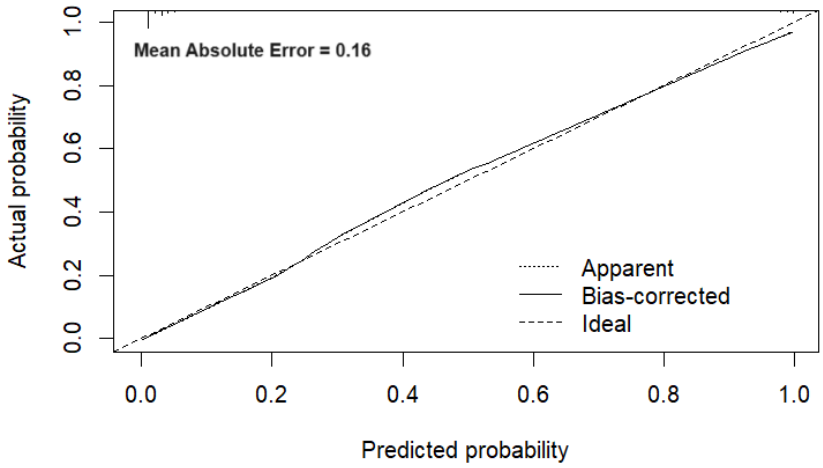

Figure 5. Calibration curve of the final trained model *The dashed line stands for the ideal calibration; the dotted line illustrates the apparent (uncorrected) calibration curve, whereas the solid line represents the bias-corrected curve derived from 1000 bootstrap samples.*

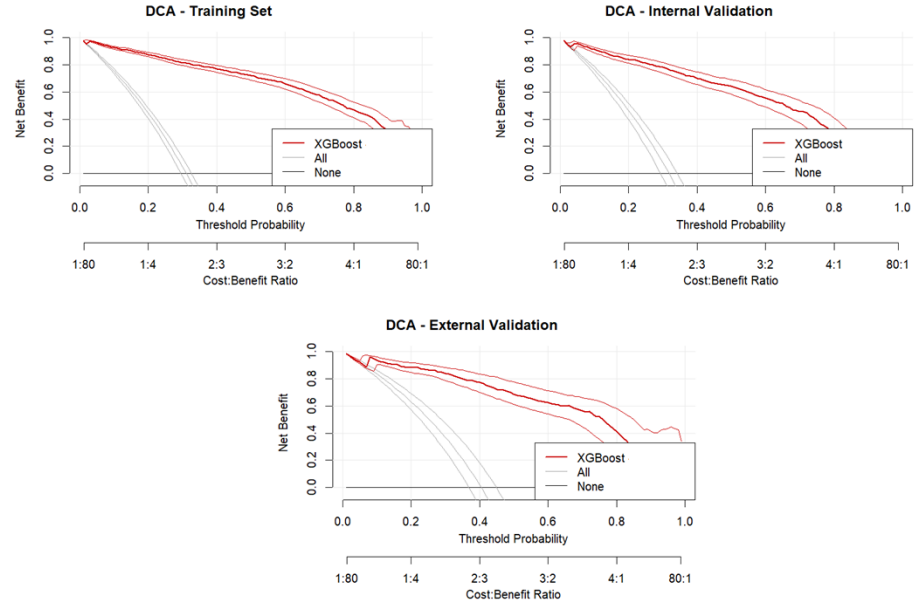

Figure 6. Decision curve analysis for the final trained model across training, internal validation, and external validation cohorts.

Table 5. Performance parameters of the final trained model.

Parameter	Value
ROC-AUC	0.7
Accuracy	0.84
Precision	0.78
F1	0.81
Recall	0.81
Brier Score	0.13

ROC-AUC: Receiver operating characteristic area under the curve.

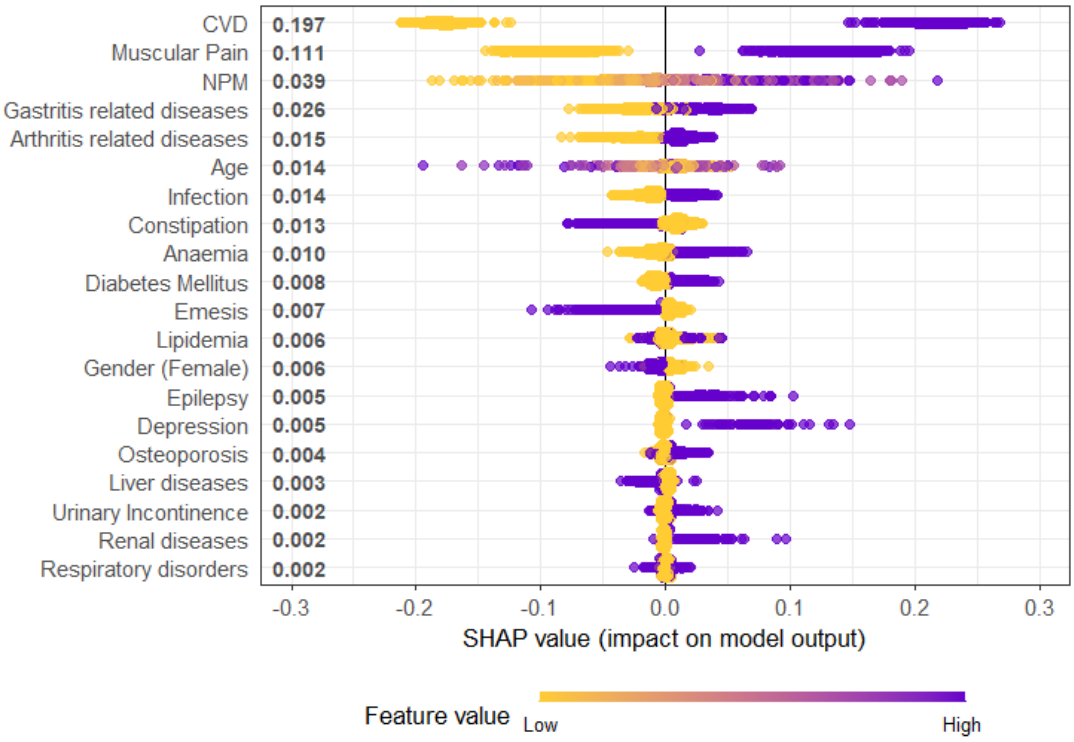

Figure 7. SHapley Additive exPlanations (SHAP) plot illustrating predictors contributions to the final model. *CVD: Cardiovascular diseases, NPM: Number of prescribed medications in emergency department.*

of 20 – 60% (**Figure 6**). The final model performance parameters are summarized in **Table 5**.

SHAP analysis revealed that the top predictors contributing to PIP risk include cardiovascular diseases, musculoskeletal pain, the number of medications prescribed in the ED, gastritis-related disorders, and arthritis-related conditions (**Figure 7**). The SHAP bee-swarm plot demonstrates the direction and intensity of each predictor's influence on specific predictions. Additionally, SHAP force charts were utilized for specific cases to provide individualized risk interpretation.

3.5. Model Deployment and Accessibility

The final trained model was deployed as an interactive web-based tool using the ShinyApp platform to enhance usability and support clinical integration. This tool enables users to enter individual patient characteristics and obtain real-time predictions of PIP risk. The platform is publicly accessible, requires no specific software or programming expertise, and is designed to facilitate real-time decision-making in emergency care settings (https://aymenali.shinyapps.io/pim_shin1/).

4. Discussion

This study trained and externally evaluated an ML model utilizing the XGBoost algorithm to predict the risk and burden of PIP among older persons presenting to the ED. The model, based on a retrospective dataset of 4,942 patients aged 65 years and older, had satisfactory discriminative performance (AUC-ROC = 0.70), high accuracy (84%), and clinically significant usefulness, as validated by DCA in the external validation set. Significant predictors encompassed cardiovascular diseases, the total number of medications prescribed in the ED, musculoskeletal pain, gastritis-related problems, and arthritis-related conditions. Moreover, SHAP plot analysis improved model interpretability by visually illustrating the contribution of each predictor to individualized risk assessments. These findings support the potential application of this model for real-time risk evaluation and personalized decision-making in acute care environments. It should be interpreted as evidence

supporting the contextual applicability of ML based PIP prediction in the ED, rather than as the development of a universally generalizable predictive model.

This study's significant accomplishment is the effective implementation of a high-performing, interpretable model utilizing readily available clinical and medication data. Previous studies examining the prevalence of PIP in older populations have predominantly concentrated on patients in primary or long-term care settings and often relied on logistic regression or rule-based algorithms to identify associated factors [9, 12, 22]. In contrast, this study focuses on the ED, a setting characterized by urgent decision-making, frequent polypharmacy, and incomplete medical histories [5]. The findings coincided with a growing body of studies endorsing the efficacy of artificial intelligence (AI) in reducing medication-related harm in older persons, who are especially vulnerable to incorrect prescribing due to altered pharmacokinetics and various chronic illnesses [11].

A significant finding of this study is the remarkably high prevalence of PIP within the population studied. More than 80% of elderly ED patients demonstrated at least one PIP, exceeding rates recorded in similar populations. Prior studies indicated that PIP occurred in 35% of older adults visiting geriatric clinics and 40% among those with dementia [13, 22]. A prevalence of 78% was observed among patients discharged from geriatric cardiovascular wards [12]. The elevated PIP rate noted in the present study may be attributed to the nature of the ED setting, where rapid assessments, polypharmacy, and time constraints increase the likelihood of suboptimal prescribing. These findings underscore an urgent need for real-time clinical decision support tools tailored to the demands of emergency care.

Cardiovascular disease (CVD) emerged as the primary predictor of PIP risk in the model. This aligns with current evidence indicating that older adults with CVD frequently present with multiple comorbidities, necessitating complex and long-term pharmacologic regimens [23, 24]. These patients are particularly susceptible to polypharmacy, a recognized risk factor for PIP, drug–drug interactions, and adverse drug reactions [25]. Moreover, several medications commonly used in the management of CVD, including diuretics, beta-blockers, antiplatelets, and anticoagulants, are often identified by STOPP criteria as potentially inappropriate, especially when prescribed without explicit indications or in the context of contraindications such as renal impairment or a history of bleeding [3]. This emphasizes the need for targeted medication review protocols for older patients with cardiovascular conditions, particularly in high-pressure environments like the ED.

Another important finding was the predictive impact of musculoskeletal pain and arthritis-related conditions on PIP risk. Pain is a prevalent issue among older persons and often necessitates pharmacologic intervention with medications that carry substantial safety concerns [26, 27]. NSAIDs are frequently used to relieve musculoskeletal and arthritic pain; however, they are deemed unsuitable for elderly people with comorbidities such heart failure, renal disease, or gastrointestinal disorders [3]. Despite these concerns, NSAIDs continue to be extensively administered, particularly in acute care settings [28, 29]. Furthermore, opioids may be prescribed without sufficient evaluation of non-pharmacologic options or the patient's functional level, hence increasing the risk of drowsiness, delirium, falls, and cognitive deterioration [30]. The relationship between pain management and PIP risk emphasizes the necessity of integrating both clinical context and patient-specific risk factors into prescribing decisions, especially in pain management.

Although polypharmacy, cardiovascular and musculoskeletal diseases were confirmed as risk factors for PIP in the SHAP analysis, these findings aligned with previous published research which linked these variables to PIP prescribing. These findings may limit variable discovery novelty; but it increases the model's clinical credibility. The final model further demonstrates how these variables' relative importance, directionality, and interaction patterns differ in the ED context, despite their individual significance. The

model demonstrates the clinical relevance of medication burden at the ED by showing that acute polypharmacy exposure during the ED encounter has a stronger marginal effect on PIP risk than several chronic disease indicators using SHAP analysis. Early ML-based PIP studies did not consistently show this ED-specific risk hierarchy.

The current ML model had ROC-AUC 0.70 and accuracy 0.84 for binary PIP presence in external validation. Recent ML-based PIM models have reported higher discrimination in clinically homogeneous cohorts, such as Yang et al.'s 0.894 external ROC-AUC in elderly stroke patients [9]. However, externally transported acute-care prediction tasks have AUCs closer to 0.77 [21]. Likewise, other ML-based PIP models created for outpatient or disease-specific inpatient cohorts have indicated AUC values between around 0.80 and 0.90, demonstrating superior discriminative efficacy under more stable prescription environments [12, 22]. Main explanation for this slightly difference could be linked to the fragmented medication histories, time-critical decision-making, and heterogeneous patient presentations may limit discrimination in the ED. Within this context, the present model prioritizes robust calibration, clinical utility, and interpretability over maximal AUC. Decision curve analysis demonstrated meaningful net benefit across clinically relevant thresholds, supporting the model's role as a risk-screening and prioritization tool to help clinicians identify older ED patients who may benefit from targeted medication review. Finally, the incremental contribution of this model is its ED-specific validation, use of routinely available data, temporal external validation, and real-world deployment, not in outperforming models developed in less time-critical care settings.

The reported accuracy of 84% may indicate robust performance; however, given the high frequency of PIP in the population, it is crucial to acknowledge the potential effects of class imbalance. To address this issue, we assessed model performance using supplementary metrics, including the AUC (0.70), sensitivity (81%), and specificity (78%), all indicating satisfactory discrimination. The results demonstrate that the model is both accurate and proficient in differentiating between PIP and non-PIP situations, maintaining a suitable balance between true positives and true negatives.

In the external validation set, a subset of patients with PIP was categorized as low risk by the model, demonstrating the intrinsic trade-off between sensitivity and specificity in predictive modelling. Although the model exhibited satisfactory discrimination overall, false negatives continue to be clinically significant, particularly in elderly patients with intricate prescription regimens.

Several strengths of this study should be highlighted. Initially, it utilized a large, real-world dataset obtained from electronic health records of a high-risk ED population, thereby augmenting the model’s applicability to clinical practice. This study specifically examined the risk of PIP in ED settings, distinguishing itself from prior studies that primarily concentrated on primary or long-term care settings. Older patients in the ED are especially susceptible to rapid medication changes, fragmented clinical information, and acute medical presentations [5, 9, 12]. The implementation of the revised STOPP/START criteria (version 3) facilitates a comprehensive and contemporary assessment of PIM in contrast to previous criteria or the widely used Beer’s criteria, which does not address under-prescribing and anticholinergic burden [3, 4]. The model not only predicts the presence of PIP but also quantifies their burden severity, offering enhanced clinical granularity. Model training employed a strict technique, utilizing LASSO for variable selection and external validation, thus strengthening the robustness and generalizability of the results. The application of SHAP-based interpretation improves transparency and promotes clinician trust in the model's predictions.

Additionally, the methodological approach employed in this study aligns with prior research on predicting PIP through a ML framework. The originality of this work is not algorithmic innovation; it is its contribution, which is predominantly contextual and implementation-oriented, illustrating the application of existing ML algorithms in the ED setting. This restricts claims of methodological innovation but enhances the

clinical significance of the findings by tackling a prescribing context marked by time constraints, polypharmacy, and insufficient clinical data, features that are less pronounced in previous inpatient or long-term care research. Numerous ML-based tools for predicting PIP have been reported as the majority were created for inpatient, outpatient, or disease-specific settings and were not intended for real-time application in ED. The current model clinically distinguishes itself by focusing on the ED, addressing both over- and under-prescribing, and offering an interpretable, implementable tool to facilitate rapid prescribing decisions.

Nonetheless, several limitations must be acknowledged. The retrospective, single-center design may restrict the applicability of the findings to other healthcare systems or populations with varying demographic or prescribing patterns. This limitation was somewhat alleviated by the incorporation of temporally separate external validation cohort. The identification of PIM was limited to structured data available in electronic health records. Consequently, important factors such as clinical reasoning, patient preferences, and undocumented indications or contraindications may not have been fully captured. Third, the model has not yet been prospectively validated or integrated into clinical workflows, limiting assessment of its real-time effectiveness and impact on prescribing behavior. Although predicting drug-specific adverse reactions may be clinically beneficial, such approach require comprehensive longitudinal data and distinctly traceable drug-event correlations, which are generally inaccessible for the current study. However, the total number of PIP has consistently been correlated with adverse drug events, hospital utilization, and mortality in older adults. In ED, where clinicians often initiate or adjust multiple medications under time constraints, cumulative prescribing load serves as a clinically significant and actionable risk indicator.

The external validation sample might be considered small cohort; however, it represents real-world data heterogeneity and was adequate for assessing model discrimination, calibration, and clinical utility. The model exhibited consistent performance in this temporally independent cohort, confirming its robustness and generalizability. Further studies need to address these deficiencies by implementing the model in varied environments and assessing its impact on clinical outcomes, including adverse drug events, hospital readmissions, and medication reconciliation rates.

This study addresses a pressing need in emergency care by providing a real-time tool for improving prescribing safety among older adults. In the ED, where clinicians frequently lack access to comprehensive medication review resources and must make rapid decisions, this model might be used as an essential clinical decision assistance instrument. Enabling the prompt identification of patients at increased risk of PIP may facilitate targeted pharmacist intervention, enhance prescribing appropriateness, and ultimately mitigate harm in those who are particularly susceptible to medication-related problems. The incorporation of the model into electronic prescribing systems or clinical dashboards may facilitate the modernization of geriatric pharmacology and pharmacotherapy and improve patient safety in acute care settings.

Finally, the implementation of ML technologies into ED practices presents significant ethical issues, especially with healthcare provider autonomy and accountability. The proposed model is designed not to replace clinical judgment or mandate automated prescribing decisions, but to serve as an advisory risk stratification tool that identifies patients who may require more thorough medication assessment. The ultimate prescribing decisions rest solely with the treating clinician, who must balance model output with clinical context, patient preferences, and expertise.

5. Conclusions

This work trained and externally evaluated an interpretable ML model for predicting PIP among elderly ED patients. The final model exhibited robust performance and practical applicability, providing a scalable tool to facilitate safer prescribing decisions in acute care. Its integration into clinical decision support

systems may improve medication safety and contribute to the optimization of geriatric pharmacotherapy.

Supplementary Materials: No supplementary materials.

Author Contributions:

Dr. Aymen A. AlQurain: Data collection and coding and preparing data for analysis, data analysis, model development, model validation, writing first draft, evaluating, modifying and approving final manuscript.

Acknowledgments:

The author wants to express gratitude to the medical records and information technology staff of Qatif Central Hospital for their continuous assistance in the review and acquisition of patient data. The author expresses gratitude to Dr. Afnan Alshanbri and Mr. Murtadha Albaharnah for their validation of the detected PIP. Additionally, I extend my gratitude to Dr. Fadhel Alomar for assessing and reviewing the final edition.

Conflicts of Interest:

The author owns no potential conflicts of interest related to the content of this manuscript.

Funding:

The author extends his appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through project number NBU-FFR2025-396-04.

Institutional Review Board Statement:

This study used a retrospective analysis of secondary data. Data were extracted after approval from the Institutional Review Board (IRB) at Mohammed Al-Mana College for Medical Sciences (SR/RP/79, dated 17/02/2022) and the IRB at QCH (QCH-SREC019/2022, dated 08/06/2022). Demographic data, comorbidities, and recent laboratories test results of patients were obtained from the electronic medical records, whereas commonly prescribed long-term drugs and those prescribed in the emergency department were sourced from the pharmacy electronic records.

Data availability:

The datasets investigated in this study will be accessible upon reasonable request. Data will be accessible for scientific reasons to researchers whose suggested utilization has received approval from the research team.

Informed Consent Statement:

Patient consent was exempted owing to the retrospective nature of the study design.

Declaration of Generative AI and AI-assisted technologies in the writing process:

During the preparation of this work the authors used ChatGPT 4.0 and Google Gemini in order to enhance the writing of this manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

1. Van Dam CS, Labuschagne HA, van Keulen K, Kramers C, Kleipool EE, Hoogendijk EO, Knol W, Nanayakkara PW, Muller M, Trappenburg MC. Polypharmacy, comorbidity and frailty: a complex interplay in older patients at the emergency department. European geriatric medicine. 2022;13(4):849-57. https://doi.org/10.1007/s41999-022-00664-y
2. Baruth JM, Gentry MT, Rummans TA, Miller DM, Burton MC. Polypharmacy in older adults: the role of the multidisciplinary team. Hosp Pract (1995). 2020;48(sup1):56-62. https://doi.org/10.1080/21548331.2019.1706995
3. O'Mahony D, Cherubini A, Guiteras AR, Denkinger M, Beuscart J-B, Onder G, Gudmundsson A, Cruz-Jentoft AJ, Knol W, Bahat G. STOPP/START criteria for potentially inappropriate prescribing in older people: version 3. European geriatric medicine. 2023;14(4):625-32. https://doi.org/10.1007/s41999-023-00777-y
4. Panel AGSBCUE, Fick DM, Semla TP, Beizer J, Brandt N, Dombrowski R, DuBeau CE, Eisenberg

W, Epplin JJ, Flanagan N. American Geriatrics Society 2015 updated beers criteria for potentially inappropriate medication use in older adults. Journal of the American Geriatrics Society. 2015;63(11):2227-46. https://doi.org/10.1111/jgs.13702
5. Mostafa R, El-Atawi K. Strategies to measure and improve emergency department performance: a review. Cureus. 2024;16(1). https://doi.org/10.7759/cureus.52879
6. Chang C-B, Lai H-Y, Hwang S-J, Yang S-Y, Wu R-S, Liu H-C, Chan D-C. Prescription of potentially inappropriate medication to older patients presenting to the emergency department: a nationally representative population study. Scientific Reports. 2018;8(1):11727. https://doi.org/10.1038/s41598-018-30184-4
7. Molist-Brunet N, Sevilla-Sánchez D, Puigoriol-Juvanteny E, Espaulella-Ferrer M, Amblàs-Novellas J, Espaulella-Panicot J. Factors associated with the detection of inappropriate prescriptions in older people: a prospective cohort. International journal of environmental research and public health. 2021;18(21):11310. https://doi.org/10.3390/ijerph182111310
8. Harrison L, O'Connor E, Jie C, Benzoni T, Renner CH, McCracken R. Potentially inappropriate medication prescribing in the elderly: Is the Beers Criteria relevant in the Emergency Department today? The American Journal of Emergency Medicine. 2019;37(9):1734-7. https://doi.org/10.1016/j.ajem.2019.05.052
9. Yang X, Ye Q, Zhang M, Xu Y, Yang M. Development and validation of a machine-learning model for the risk of potentially inappropriate medications in elderly stroke patients. Frontiers in Pharmacology. 2025;16:1565420. https://doi.org/10.3389/fphar.2025.1565420
10. Sharma V, Kulkarni V, Jess E, Gilani F, Eurich D, Simpson SH, Voaklander D, Semenchuk M, London C, Samanani S. Development and validation of a machine learning model to estimate risk of adverse outcomes within 30 days of opioid dispensation. JAMA Netw Open. 2022;5(12):e2248559-e. https://doi.org/10.1001/jamanetworkopen.2022.48559
11. Wilhelm C, Steckelberg A, Rebitschek FG. Benefits and harms associated with the use of AI-related algorithmic decision-making systems by healthcare professionals: a systematic review. The Lancet Regional Health–Europe. 2025;48. https://doi.org/10.1016/j.lanepe.2024.101145
12. Xingwei W, Huan C, Mengting L, Lv Q, Jiaying Z, Enwu L, Jiuqun Z, Rongsheng T. A machine learning-based risk warning platform for potentially inappropriate prescriptions for elderly patients with cardiovascular disease. Frontiers in Pharmacology. 2022;13:804566. https://doi.org/10.3389/fphar.2022.804566
13. Hu Q, Tian F, Jin Z, Lin G, Teng F, Xu T. Developing a warning model of potentially inappropriate medications in older Chinese outpatients in tertiary hospitals: a machine-learning study. Journal of Clinical Medicine. 2023;12(7):2619. https://doi.org/10.3390/jcm12072619
14. Widagdo I, Kerr M, Kalisch Ellett L, Schlegel C, Sadeqzadeh E, Wang A, Clarke AL, Pratt N. Validity of the Updated Rx-Risk Index as a Disease Identification and Risk-Adjustment Tool for Use in Observational Health Studies. Clinical Interventions in Aging. 2025:309-23. https://doi.org/10.2147/CIA.S494145
15. World Health Organisation. International Statistical Classification of Disease and Related Health Problems 10th Revision (ICD-10)-WHO Version for 2016. Geneva: World Health Organization 2016.
16. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373-83. https://doi.org/10.1016/0021-9681(87)90171-8
17. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron. 1976;16(1):31-41. https://doi.org/10.1159/000180580
18. World Health Organization. Anatomical therapeutic chemical (ATC) classification index with defined daily doses (DDDs). In: Methodology CCfDS, editor. Oslo: World Health Orgnisation; 2000.
19. Grazal CF, Anderson AB, Booth GJ, Geiger PG, Forsberg JA, Balazs GC. A machine-learning algorithm to predict the likelihood of prolonged opioid use following arthroscopic hip surgery. Arthroscopy: The Journal of Arthroscopic & Related Surgery. 2022;38(3):839-47. e2. https://doi.org/10.1016/j.arthro.2021.08.009
20. Lin H-C, Wang Z, Hu Y-H, Simon K, Buu A. Characteristics of statewide prescription drug monitoring programs and potentially inappropriate opioid prescribing to patients with non-cancer chronic

pain: A machine learning application. Preventive medicine. 2022;161:107116. https://doi.org/10.1016/j.ypmed.2022.107116
21. Guan C, Gong A, Zhao Y, Yin C, Geng L, Liu L, Yang X, Lu J, Xiao B. Interpretable machine learning model for new-onset atrial fibrillation prediction in critically ill patients: a multi-center study. J Crit Care. 2024;28(1):349. https://doi.org/10.1186/s13054-024-05138-0
22. Hu Q, Zhao M, Teng F, Lin G, Jin Z, Xu T. A model for identifying potentially inappropriate medication used in older people with dementia: a machine learning study. International Journal of Clinical Pharmacy. 2024;46(4):937-46. https://doi.org/10.1007/s11096-024-01730-0
23. Villén N, Roso-Llorach A, Gallego-Moll C, Danes-Castells M, Fernández-Bertolin S, Troncoso-Mariño A, Monteagudo M, Amado E, Violán C. Polypharmacy patterns in multimorbid older people with cardiovascular disease: longitudinal study. Geriatrics. 2022;7(6):141. https://doi.org/10.3390/geriatrics7060141
24. Tamargo J, Kjeldsen KP, Delpón E, Semb AG, Cerbai E, Dobrev D, Savarese G, Sulzgruber P, Rosano G, Borghi C. Facing the challenge of polypharmacy when prescribing for older people with cardiovascular disease. A review by the European Society of Cardiology Working Group on Cardiovascular Pharmacotherapy. European Heart Journal-Cardiovascular Pharmacotherapy. 2022;8(4):406-19. https://doi.org/10.1093/ehjcvp/pvac005
25. Rodrigues MCS, de Oliveira C. Drug-drug interactions and adverse drug reactions in polypharmacy among older adults: an integrative review. Revista Latino-Americana De Enfermagem. 2016;24:e2800. https://doi.org/10.1590/1518-8345.1316.2800
26. Tracy B, Morrison RS. Pain management in older adults. Clinical therapeutics. 2013;35(11):1659-68. https://doi.org/10.1016/j.clinthera.2013.09.026
27. Al-Qurain AA, Gebremichael LG, Khan MS, Williams DB, Mackenzie L, Phillips C, Russell P, Roberts MS, Wiese MD. Prevalence and Factors Associated with Analgesic Prescribing in Poly-Medicated Elderly Patients. Drugs Aging. 2020;37(4):291-300. https://doi.org/10.1007/s40266-019-00742-0
28. Gnjidic D, Blyth FM, Le Couteur DG, Cumming RG, McLachlan AJ, Handelsman DJ, Seibel M, Waite L, Naganathan V. Nonsteroidal anti-inflammatory drugs (NSAIDs) in older people: prescribing patterns according to pain prevalence and adherence to clinical guidelines. PAIN®. 2014;155(9):1814-20. https://doi.org/10.1016/j.pain.2014.06.009
29. Alqurain AA, Albaharnah M, Al Zayer S, Ameer L, Ghosn S, Al-Shaibi S, Algoraini M, Aldhafeeri A, Alyusuf DA, Alshnbari A, et al. The prevalence of polypharmacy and hyper-polypharmacy among middle-aged vs. older patients in Saudi Arabia: a cross-sectional study. Front Pharmacol. 2024;15:1357171. https://doi.org/10.3389/fphar.2024.1357171
30. Al-Qurain AA, Gebremichael LG, Khan MS, Williams DB, Mackenzie L, Phillips C, Russell P, Roberts MS, Wiese MD. Opioid prescribing and risk of drug-opioid interactions in older discharged patients with polypharmacy in Australia. International Journal of Clinical Pharmacy. 2020. https://doi.org/10.1007/s11096-020-01191-1