Using machine learning to predict pharmaceutical interventions during medication prescription review in a hospital setting

Erin Johns, PhD, Direction de la Qualité, de la Performance et de l'Innovation, Agence Régionale de Santé Grand Est, Strasbourg, France; ICube – IMAGeS, UMR 7357, Université de Strasbourg, France, and Laboratoire de Pharmacologie et de Toxicologie Neurocardiovasculaire, UR7296, Faculté de médecine, Strasbourg, France

Ahmed Guendouz, MD, Direction de la Qualité, de la Performance et de l'Innovation, Agence Régionale de Santé Grand Est, Strasbourg, France, and Laboratoire de Pharmacologie et de Toxicologie Neurocardiovasculaire, UR7296, Faculté de médecine, Strasbourg, France *

Laurent Dal Mas, MSc, Direction de la Qualité, de la Performance et de l'Innovation, Agence Régionale de Santé Grand Est, Strasbourg, France

Morgane Beck, PhD, Direction de la Qualité, de la Performance et de l'Innovation, Agence Régionale de Santé Grand Est, Strasbourg, France

Ahmad Alkanj, PhD, Laboratoire de Pharmacologie et de Toxicologie Neurocardiovasculaire, UR7296, Faculté de Médecine, Strasbourg, France

Bénédicte Gourieux, PharmD, Laboratoire de Pharmacologie et de Toxicologie Neurocardiovasculaire, UR7296, Faculté de Médecine, Strasbourg, France and Service Pharmacie-Stérilisation, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Erik-André Sauleau, PhD, ICube – IMAGeS, UMR 7357, Université de Strasbourg, France, and Département de Santé Publique – Groupe Méthodes Recherche Clinique, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Bruno Michel, PhD, Laboratoire de Pharmacologie et de Toxicologie Neurocardiovasculaire, UR7296, Faculté de Médecine, Strasbourg, France, and Service Pharmacie-Stérilisation, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Address correspondence to Dr. Guendouz (ahmed.guendouz@ars.sante.fr).

https://doi.org/10.1093/ajhp/zxaf089

Objective: Medication errors are a worldwide public health issue. Reducing inappropriate medication use is a daily challenge for clinical pharmacists. Computerization of the medication process and the rise of artificial intelligence make it possible to develop tools to detect inappropriate prescriptions. Our main goal was to compare the performance of two machine learning models capable of predicting the probability of a prescription requiring pharmaceutical intervention (PI) using hospital data.

Methods: The study was conducted in a single hospital, with data collected over 4 years, including 2,059,847 prescription lines (a patient's entire medication regimen consists of multiple prescription lines) associated with 260,611 PIs. Two tree-based binary classification machine learning models were tested: the Light Gradient Boosting Machine (LGBM) model and the Random Forest (RF) model. The dataset was split (70% for training and 30% for testing), and training and testing were performed on the global dataset and on data stratified by medical care department.

Results: For the global dataset, the LGBM model outperformed the RF model in most metrics: accuracy (86% vs 85%), precision (80% vs 42%), specificity (97% vs 89%), area under the curve (83% vs 71%) and F1-score (58% vs 47%). However, the RF model had superior recall (53% vs 46%). Furthermore, the LGBM model trained on the global database was generally more effective than models trained on the care departments' databases.

Conclusion: The LGBM model showed superior performance in detecting inappropriate prescriptions, potentially improving the thoroughness and efficiency of prescription review. While further studies are needed to confirm these findings, the model holds significant promise for advancing hospital clinical pharmacy and enhancing patient care through optimized prescription management.

Keywords: artificial intelligence, clinical pharmacy, high-risk prescriptions, inappropriate prescribing, machine learning

Am J Health-Syst Pharm. 2025;82:1238-1248

Medication errors are a worldwide issue. According to the US National Coordinating Council for Medication Error Reporting and Prevention,[1] a medication error is "any preventable event that may cause or lead to inappropriate medication use or patient harm while the medication is in the control of the healthcare professional, patient, or consumer." It can result from a wrong indication for a medication, an incorrect dose or treatment duration, drug interactions, failure to initiate a medication, or initiation of a medication when not appropriate in a specific context. In 2017, the World Health Organization initiated the Medication Without Harm program[2] to reduce severe avoidable harm related to medication use by 50% in 5 years. Worldwide, medication-related harm is preventable in 50% of cases, and the annual cost of

medication errors is estimated at $42 billion.[3]

Clinical pharmacists have a predominant role in limiting errors through medication review. Pharmaceutical Care Network Europe (PCNE)[4] has defined medication review as follows: "structured evaluation of a patient's medicines with the aim of optimizing medicines use and improving health outcomes. This entails detecting drug-related problems and recommending interventions." A pharmaceutical intervention (PI) is defined as "any activity undertaken by the pharmacist which benefits the patient."[5] PIs are meant to prevent negative outcomes and optimize therapy when a medication error is detected. Several studies[6-8] show that PIs have positive clinical, economic, and organizational impacts.

Today, the computerization of the medication-use process in hospitals and the development of clinical decision support (CDS) systems help hospital pharmacists in the medication review process. While well-designed CDS systems can offer many benefits, such as improving efficiency and decision-making, they may also lead to alert fatigue when nonrelevant or inappropriate alerts are generated. This can reduce the effectiveness of the system.[9] However, computerization of pharmaceutical activities has led to the collection of a massive quantity of data. The rise of artificial intelligence (AI) in pharmacy is an opportunity to leverage this data, reduce "noise" by filtering out irrelevant alerts, and assist clinical pharmacists in prioritizing high-impact interventions.[10,11,12].

AI, thanks to machine learning (ML), has the capacity to predict situations using retrospective data. ML is a subset of AI used for addressing classification, regression, clustering, dimension reduction, or association tasks. ML models determine the rules to solve these tasks, thanks to the training dataset. The training dataset is a retrospective dataset composed of data relative to the prediction task. The training can be supervised, unsupervised, semisupervised, self-supervised, or reinforced. Supervised ML means the outcome to predict is labeled in the training dataset and is specified to the algorithm to facilitate the development of the predictive model. Supervised ML is finding its way into clinical pharmacy as a method to assist pharmacists in their activities, such as predicting adverse drug events in older inpatients to enhance medication safety,[13] identifying high-risk QTc prolongation related to drug-drug interactions,[14] and reducing medication-related risks[15] .

Objectives. The objectives of this study were 2-fold. The first objective was to compare the performance of 2 common supervised ML methods in order to generate algorithms (from the same dataset) capable of predicting the probability that a prescription requires PIs in a hospital setting. The performance comparison between the models was conducted using 6 metrics: accuracy, recall, precision, specificity, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Based on the initial results, the second study objective was to develop additional algorithms using selected databases targeting specific medical care departments with the aim of improving the performance of the models.

Methods

Setting. This retrospective study was conducted at the University Hospital of Strasbourg in France. This hospital offers 1,972 beds for medical, surgical, and obstetrics activities grouped in more than 20 care departments. The hospital manages a high patient volume, with 451 emergency visits per day (164,575 annually) and approximately 46,741 hospital stays each year. The average length of stay is 5.9 days. In all care units except for medical and surgical intensive care units, the patients' prescription lines (a patient's entire medication regimen consists of multiple prescription lines) are prescribed using the prescription assistance software DxCare (Dedalus France, Artigues-près-Bordeaux, France), while biological orders are filed in the Clinysys GLIMS software (Clinisys, Inc., Tucson, AZ). Clinical pharmacists perform their medication review activity on DxCare and notify the clinicians if a prescription requires a PI through a brief comment explaining the drug-related issue and suggesting an appropriate prescription modification.

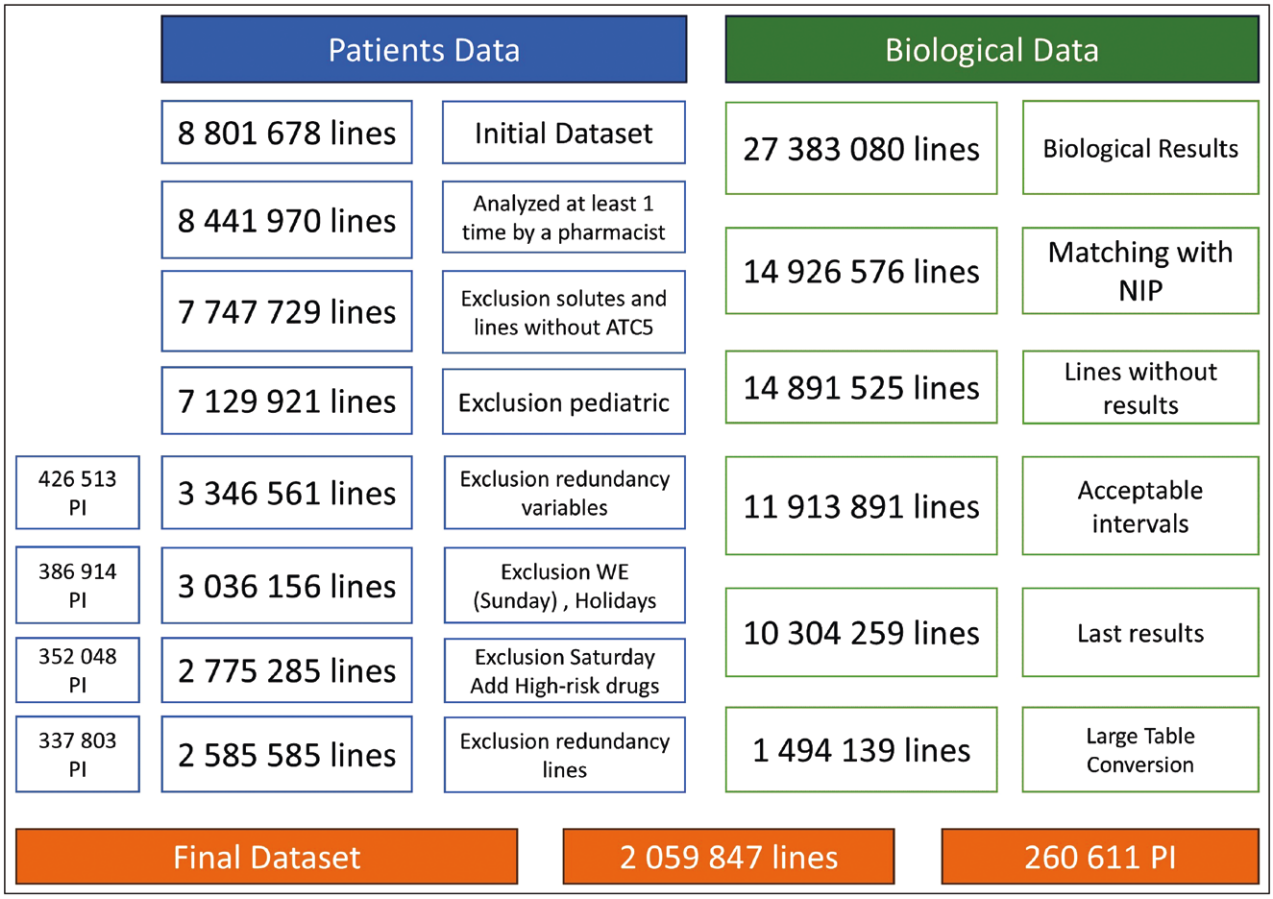
Dataset. ***Data collection.*** Data collected over a 4-year period (2017-2020) were extracted from the electronic health record. This study covered 97,842 patients hospitalized across all care units using the DxCare software. Prescription lines, PIs generated by the clinical pharmacists, biological results, and hospitalization and administrative data for all inpatients were collected to train the ML models. After data cleaning and processing (Figure 1), the final dataset consisted of 2,059,847 prescription lines associated with 260,611 PIs.

Data preparation. Before training a model, the extracted data was preprocessed. After a first analysis of the data, outliers and duplicates were deleted. Clinically irrelevant PIs, such as those related to drugs not listed in the hospital's formulary, were also removed. To handle missing biological values, the main interest was whether the data was present or absent and, if present, whether it was within the standard range or not. To address this, biological results were dichotomized in 4 categories to bypass the large number of missing values: 0 (missing value), 1 (value below the standard), 2 (value in the standard range), 3 (value above the standard). Finally, prescription lines were also dichotomized (0 = no PI, 1 = presence of PI) to simplify the handling of the text data.

To satisfy the second objective, the extracted and preprocessed data was divided into datasets for 9 selected care departments: ophthalmology, geriatric, cardiovascular pathologies, thoracic pathologies, head and neck pathologies, internal medicine, traumatology, emergency, and digestive pathologies and transplantation.

Model development. ***Training and test sets.*** The ML models were trained on the same dataset, covering

Figure 1. Representation of the steps involved in processing of raw data and presentation of the remaining data after preprocessing and removal of outlier values. ATC indicates Anatomical Therapeutic Chemical; NIP, patient identification number; WE, weekend.



all of the care departments. The overall dataset was randomly split in 2 sets, with 70% of the dataset used to train the models and the remaining 30% of the dataset used to test the performance of the predictive model on untrained data. This 70/30 split approach[16] was used to ensure a sufficient amount of data for both model training and evaluation.

The imbalance of the PIs in this dataset biased the training of the predictive models. Due to this imbalance, we oversampled the prescription lines requiring a PI (randomly duplicating instances from the minority class to increase their representation) and undersampled the prescription lines not associated with a PI (randomly removing instances from the majority class to balance the dataset) in the training, resulting in an equalized dataset.[17] The balancing was performed using the ovun.sample function from the ROSE package in R (R Foundation for Statistical Computing, Vienna, Austria). This approach ensures that the model learns equally from both types of prescription lines, improving its ability to detect PIs. The training set then included 1,441,892 prescription lines, of which 720,745 involved PIs. The test set was not sampled and included the remaining 30% of the original dataset: 617,955 prescription lines, of which 78,282 were associated with PIs.

The PIs (a binary 0/1 categorization) were the labels used for the training of the models, indicating the outcome to be predicted. The remaining variables (age, sex, care department code, name of the prescribed drug, fifth-level Anatomical Therapeutic Chemical [ATC] Classification System code [ATC5] for the prescribed drug, route of administration, and biological results [levels of creatinine, C-reactive protein, hemoglobin, leukocytes, potassium, platelets, and sodium, as well as international normalized ratio] associated with the prescription were the predictors necessary to develop and train the different models.

Then, the performance of the different models was compared. The ML model with the best performance was then used to train the ensuing 9 models for the selected care departments. The models for selected care departments were built on the same basis: 70% of the dataset was used for the training set and the remaining 30% for the test set. The same process of equalization used for the overall dataset was performed to

Table 1. Descriptive Statistics for Patient Data

	PIs (N = 260,611)	No PIs (N = 1,799,236)
Sex, No. (%)		
Male	140,031 (53.7)	952,702 (53.0)
Female	120,577 (46.3)	846,493 (47.0)
Unknown	3 (<0.001)	41 (<0.001)
Age, years		
Mean (SD)	64.3 (17.6)	67.8 (16.9)
Median (range)	67.0 (19.0-108)	70.0 (19.0-108)
Creatinine (mg/dL)		
Mean (SD)	1.09 (0.68)	1.08 (0.68)
Median (range)	0.95 (0.34-5.09)	0.93 (0.34-5.09)
Missing data	73,895 (28.4)	416,767 (23.2)
C-reactive protein, mg/L		
Mean (SD)	67.0 (77.5)	61.7 (74.0)
Median (range)	35.4 (4.0-450)	31.0 (4.0-450)
Missing data	122,602 (47.0)	717,932 (39,9)
International normalized ratio		
Mean (SD)	1.34 (0.633)	1.45 (0.792)
Median (range)	1.14 (0.95-9.69)	1.16 (0.95-9.97)
Missing data	131,687 (50.5)	926,974 (51,5)
Hemoglobin, g/dL		
Mean (SD)	11.2 (2.06)	11.3 (2.02)
Median (range)	11.1 (3.4-15.9)	11.2 (3.1-15.9)
Missing data	72,543 (27.8)	455,299 (25.3)
Leukocytes, g/dL		
Mean (SD)	9.66 (4.70)	9.40 (4.44)
Median (range)	8.85 (0.21-30.0)	8.58 (0.21-30.0)
Missing data	73,473 (28.2)	446,171 (24.8)
Potassium (mEq/L)		
Mean (SD)	4.01 (0.511)	3.99 (0.521)
Median (range)	3.97 (2.7-7.9)	3.94 (2.7-7.9)
Missing data	63,127 (24.2)	360,210 (20.0)
Platelets, $\times 10^9$/L		
Mean (SD)	244 (111)	248 (108)
Median (range)	234 (11.0-599)	236 (11.0-599)
Missing data	75,269 (28.8)	465,704 (25.9)
Sodium, mEq/L		
Mean (SD)	138 (4.39)	138 (4.40)
Median (range)	138 (121-159)	138 (121-159)
Missing data	65,278 (25.0)	377,991 (21.0)

Abbreviation: PI, pharmaceutical intervention.

counter the imbalance of the training dataset.

ML models used. Two tree-based models used for binary classification machine learning models were tested on the training dataset: a Random Forest[18] (RF) model and a gradient boosting[19] model (Light Gradient Boosting Machine[20] [LGBM]).

A grid search was performed on each training dataset to tune and optimize the hyperparameters used to train the models. For the RF models, the following hyperparameters were calculated: number of trees and number of randomly drawn variables (mtry). As for the LGBM models, minimal node size, tree depth, and number of leaves hyperparameters were determined.

The different models were trained using R version 4.3.0. The RF and LGBM models were trained using, respectively, the randomforest (4.6-14) and lgbm (3.3.5) packages in R.

Model evaluation. ***Performance assessment metrics.*** The models' performance assessment was carried out by measuring 6 key metrics that are used in medical classification problems, particularly in the evaluation of ML models for imbalanced datasets[21,22]:

- Accuracy—the ratio of correctly classified instances (true positives and true negatives) to the total number of instances in the evaluation set, which quantifies the overall correctness of the model's predictions
- Recall (sensitivity or true positive rate)—the ability of the model to correctly identify positive instances (in our case, PIs) out of all actual positive instances
- Precision (positive predictive value)—the accuracy of positive predictions made by the model (in our case, the ratio of true-positive PIs to the total number of positive PI predictions), which quantifies how well the model can capture all PIs
- Specificity (true negative rate)—the ability of the model to correctly identify negative instances (non-PI prescriptions) out of all actual negative instances, which quantifies how well the model can avoid false-positive PIs
- F1-score—a combined metric that balances precision and recall. It provides a single score that considers all positive predictions. The F1-score is particularly useful when dealing with imbalanced datasets, where one class significantly outnumbers the other
- AUC-ROC—a score measuring the classification performance using the relationship between sensitivity and specificity

The selection of these metrics was based on best practices in medical and ML research, as they collectively provide a comprehensive evaluation of the models' performance. While accuracy is a general indicator, recall and specificity are crucial in a clinical context where false negatives (missed PIs) and false positives (unnecessary alerts) must be carefully balanced. The F1-score is particularly relevant given the imbalanced nature of the dataset, and AUC-ROC helps assess the overall discriminative ability of the models.

Statistical analysis. To compare the AUC-ROC values of the different models, we used the DeLong test,[23] a nonparametric statistical test used to compare the AUC-ROC values between different models. It assesses whether the difference in AUC-ROC between 2 models is statistically significant. A P value of <0.05 indicates that the difference is statistically significant.

Additionally, we applied the Youden index[24] to assess the maximum potential effectiveness of the predictive models. This index helps determine the optimal decision threshold by maximizing the sum of sensitivity and specificity. A model is considered more effective when its Youden index is close to 1.

Ethics approval. The local ethics committee approved this noninterventional and retrospective study (reference CE-2022-21).

Results

Datasets characteristics. From January 2017 through December 2020, a total of 2,059,847 prescription lines were reviewed by the clinical pharmacists. Of these, 260,611 prescription lines (12.7%) required a PI. As shown in Table 1, the demographic distribution of patients with PIs was compared to that of those without PIs. The majority of prescription lines for both groups were for male patients (53.7% and 53%, respectively). The median age of patients with a PI was 67 years. For the biological data variables, the missing data rate ranged from 24.2% (for potassium values) to 50.5% (for international normalized ratio values). Table 2 provides the distribution of prescription lines requiring PIs across different care departments. The percentage of prescription lines requiring PIs varied by department, with some showing higher or lower rates than the overall rate of 12.7%. The traumatology department had the highest percentage of prescription lines with PIs (16.9%), well above the global average, followed by head and neck pathologies (13.8%). In contrast, the geriatric (7.3%) and ophthalmology (9.7%) departments had lower-than-average percentages.

Comparative performance of both models on the overall test dataset. The LGBM model showed better performance than the RF model (see Table 3 and Figure 2),. outperforming it in terms of accuracy, precision, specificity, F1-score, and AUC-ROC. However, it did not show superior performance for recall. This was confirmed by the DeLong test, which shows a statistically significant difference between the AUC-ROC values (P = 0.002). It is important to note that these results are based on the oversampled test data, as performance metrics may vary when reported for raw versus oversampled data.

To enhance the performance of the LGBM model, we determined the optimal cut-off point thanks to the Youden index. As shown in Figure 3, the best cut-off point was 0.43, giving a Youden index of 0.67, meaning that 67% of the predictions were not random, with a sensitivity of 98% and a specificity of 95%.

Table 2. Distribution of Prescriptions Requiring a Pharmaceutical Intervention in Selected Care Departments

Care department	Prescriptions with PI, No. (%)	Prescriptions without PI, No. (%)
Ophthalmology	25,250 (9.7)	235,717 (90.3)
Geriatric	13,130 (7.3)	167,699 (92.7)
Cardiovascular pathologies	28,093 (10.2)	247,444 (89.8)
Thoracic pathologies	14,875 (9.3)	145,529 (90.7)
Head and neck pathologies	27,510 (13.8)	171,553 (86.2)
Traumatology	30,049 (16.9)	147,195 (83.1)
Internal medicine	37,584 (10.6)	318,435 (89.4)
Emergency	27,869 (10.5)	237,691 (89.5)
Digestive pathologies and transplantation	32,612 (12.1)	235,984 (87.9)

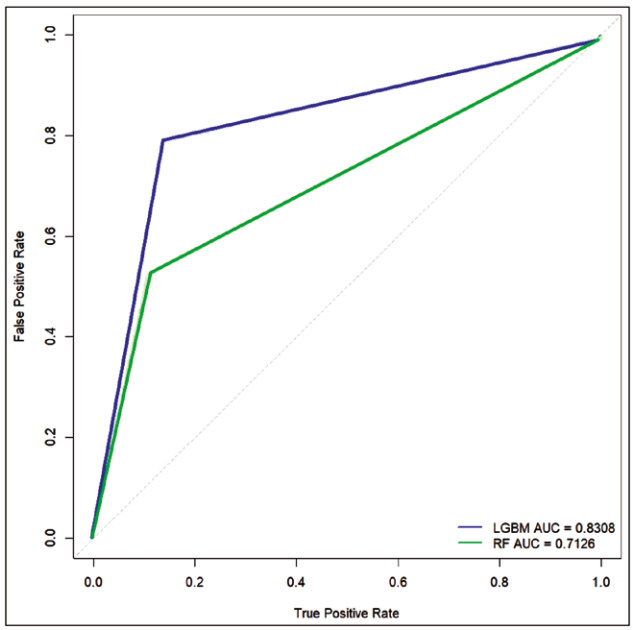
Abbreviation: PI, pharmaceutical intervention.

Table 3. Models' Performance on Testing Dataset (N = 617,955 Prescriptions, 78,282 PIs)[a]

Model	Accuracy	Recall	Precision	Specificity	F1-score	AUC-ROC
LGBM	86	46	80	97	58	83
RF	85	53	42	89	47	71

Abbreviations: AUC-ROC, area under the receiver operating characteristic curve; LGBM, Light Gradient Boosting Machine; RF, Random Forest.
[a]All data are percentages.

Figure 2. Comparison of the receiver operating characteristic (ROC) curves for the Light Gradient Boosting Machine (LGBM) model (blue curve) and the Random Forest (RF) model (green curve). True positive rate = recall (%); false positive rate = 1 – specificity.



Comparative variable importance of both models on the overall test dataset. The analysis of the variables' importance showed that predictors did not have the same importance in the models' training. The higher the score, the higher the importance of the predictor in the predictive model. However, 9 of the 10 most important features were common to the LGBM and RF models (Figures 4 and 5). These included ATC5, care unit, route of administration, age, active substance, international normalized ratio, and values for creatinine, potassium, and leukocytes.

Performance of the LGBM model (overall test dataset vs subdivided test datasets). Since the LGBM model performed better than the RF model, we chose to pursue our study with it.

Statistics for the subdivided dataset performance are presented in Table 4.

Figure 3. Plot of the optimal cut-off point for the Light Gradient Boosting Machine model.

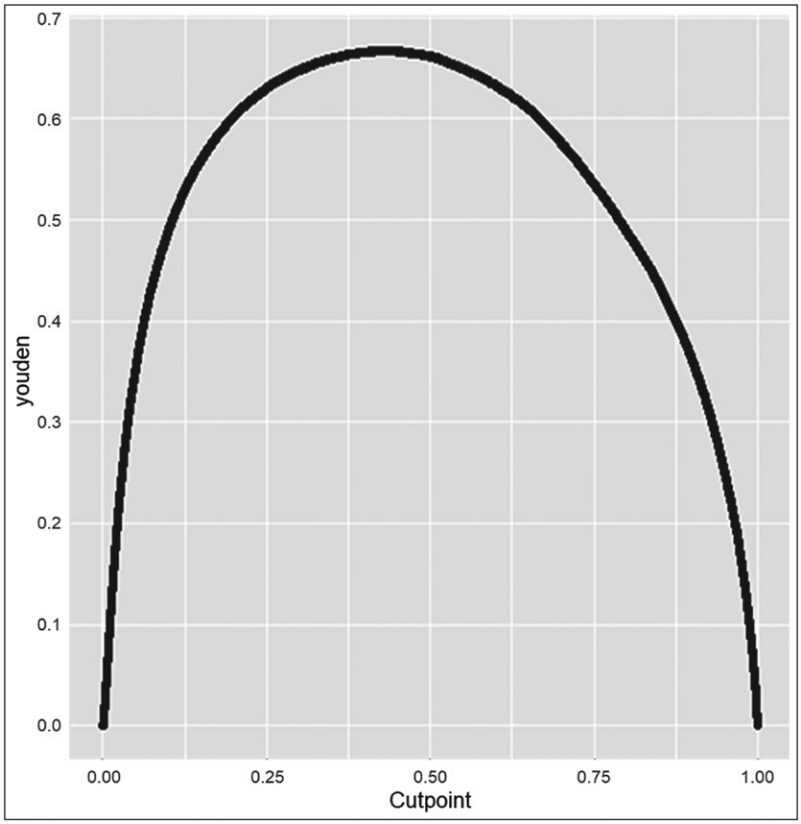
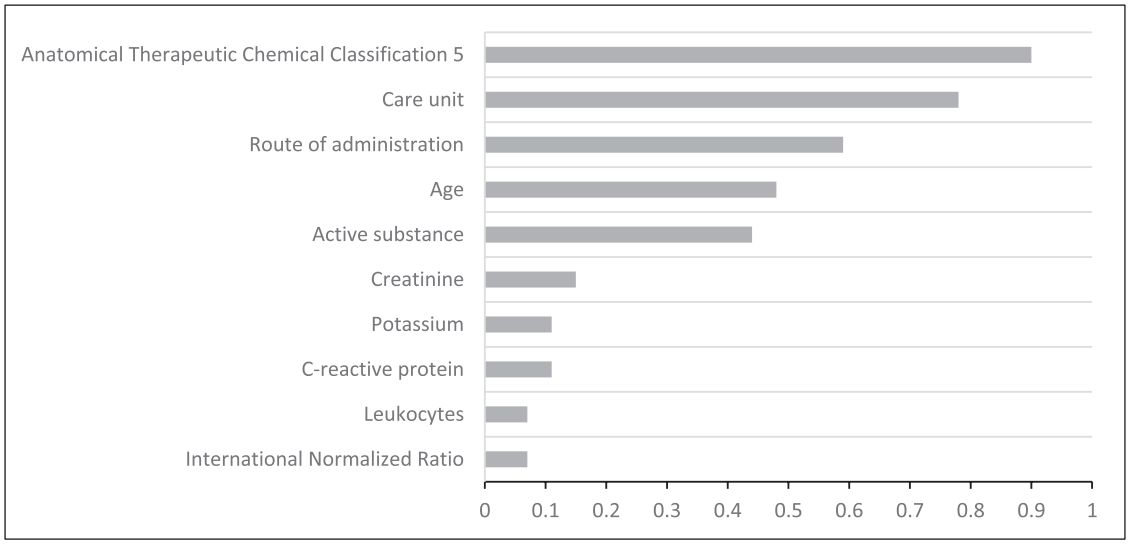


Figure 4. Relative importance of evaluated variables in predicting pharmaceutical interventions with the Light Gradient Boosting Machine model.



The results show that the models trained on the total database were, in most cases, more effective than the models trained on the care departments' databases.

Discussion

The results presented above on the overall dataset demonstrated the superiority of the LGBM model when basing model performance on F1-score and AUC-ROC values. Indeed, the results obtained with the LGBM exhibited the highest F1-score and AUC-ROC

Figure 5. Relative importance of evaluated variables in predicting pharmaceutical interventions with the Random Forest model.

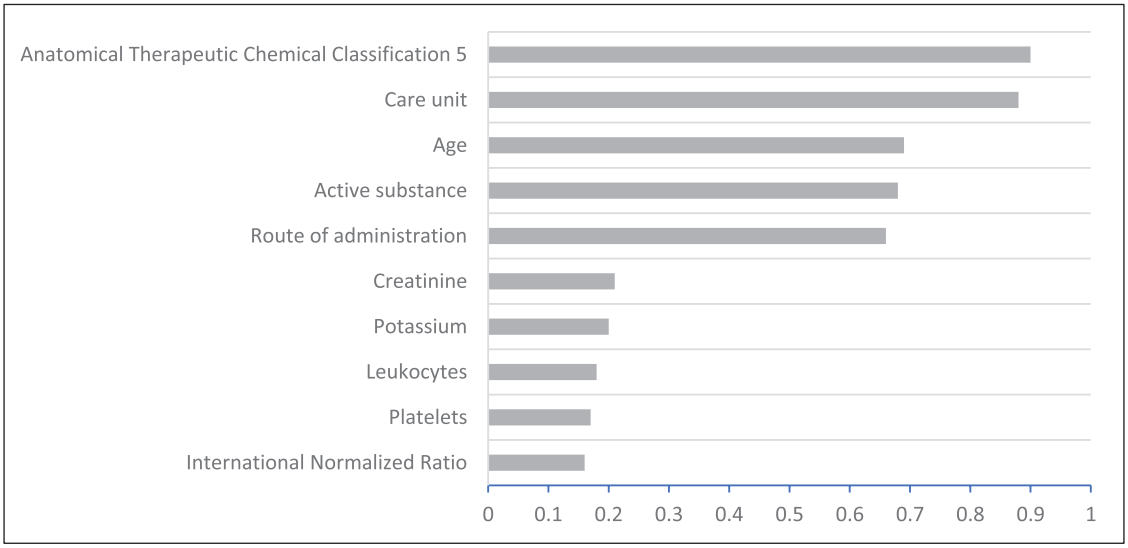


Table 4. LGBM Model's Performance on Overall Dataset and on Selected Care Departments' Testing Datasets[a]

	Accuracy	Recall	Precision	Specificity	F1-score	AUC-ROC
Overall dataset	86	46	80	97	58	83
Care department						
Ophthalmology	86	41	74	96	53	81
Geriatric	84	30	73	97	43	79
Cardiovascular pathologies	90	53	86	98	66	88
Thoracic pathologies	86	41	76	97	53	82
Head and neck pathologies	87	56	83	96	67	85
Traumatology	80	50	80	94	62	80
Internal medicine	88	49	73	96	57	81
Emergency	84	42	79	97	55	82
Digestive pathologies and transplantation	88	54	82	97	65	85

Abbreviation : AUC-ROC, area under the receiver operating characteristic curve.
[a]All data are percentages.

values, which were statistically different from RF model values according to the DeLong test.

Going into more detail, the variable importance plots show that both models gave the greatest importance to the variables directly linked to the drug prescription (ATC5, care unit, route of administration, age, and active substance). PIs are commonly linked to an improper dosage prescription or route of administration,[25] to being elderly,[26] or to high-risk medication.[27] Then come the biological results that complete the medication review.[28] This division can be explained by the preprocessing choice of the biological data and the proportion of missing values, which made it more difficult for the model to establish classification rules.

When making predictions on an imbalanced dataset, it is recommended to assess algorithm performance based on the F1-score,[29] as it does not consider negative predictions. This means that the results emphasize the algorithm's ability to maximize positive predictions. In the healthcare field, this makes sense. It is essential to focus on true positives rather than true negatives, meaning we aim to highlight the correctly predicted positive class. In our study, false positives were rare (specificity, 97%); however, false negatives should be improved (recall, 46%). The improvement will allow reduction

of false negatives and diminish fatigue due to overalerting.

Prior studies using supervised ML to detect prescription orders requiring a PI have been published.[30-35] Nonetheless, our study stands out from the literature by being the first to develop ML predictive models to detect inappropriate prescriptions on a large dataset extracted for a 4-year period and based on PIs formulated by clinical pharmacists. The lack of results-metric standardization, the volume of the databases, and the clinical issues of the published studies make it hard to compare our results with the literature. Most of the relevant previously published studies focused on a specific issue (prediction of adverse drug reactions due to vancomycin,[31] for example), and model development was based on a small dataset and/or data extracted from a short period of study. However, Hu et al[32] and Van Laere et al[14] developed models based on answering a global issue: prediction of adverse drug events among older inpatients and risk prediction of QTc prolongation. Hu et al developed a gradient boosting model that had a lower performance than our model: an F1-score of 53%, compared to a score of 58% for our LGBM model. For their part, Van Laere et al developed a gradient boosting model and an RF model. Both of our models performed better than those of Van Laere et al on the metrics of accuracy and specificity metrics: for the LGBM models, 86% vs 82% and 97% vs 87%, respectively; and for the RF models, 85% vs 82% and 89% vs 88%, respectively. For both model types, recall was lower in our study than in the study of Van Laere et al (LGBM 46% vs 73%; RF 53% vs 76%).

Thus, although our LGBM model showed lower performance in terms of recall, our goal of maximizing positive predictions led to superior specificity and accuracy, which are crucial aspects for reducing false positives and improving confidence in the model's predictions. Compared to the works we cited above, our model makes a significant contribution by better balancing different performance metrics, making it more suitable for practical clinical application.

Working with a global model can be a real asset for clinical pharmacists and assist them in their medication reviews. The model will point out inappropriate situations and accelerate the medication review process. Clinical pharmacists will then have more time to study atypical prescription orders and share their expertise with other clinicians to secure improved drug management. ML is a technology capable of adapting to its environment, which is necessary in healthcare. Reinforcement learning,[36] for example, enables ML models to update themselves by interacting with their environment using a reward/punishment system. In this way, clinical pharmacists will be able to evaluate the PI alerts from the algorithms to enhance their relevance and update the models in accordance with new scientific advancements.

While the creation of a general predictive model met our initial objective, we felt it was important to broaden the scope of investigation by developing specific models based on samples from the overall database. To the best of our knowledge, this is the first study to have developed several algorithms for detecting inappropriate hospital prescriptions based on specific care departments. Training specific models seemed to be relevant to us, particularly for the implementation of reinforced learning, which could have led to having specialized algorithms for each care department. However, our results showed that a model trained on the overall dataset globally performed better than models trained on specific datasets.

Although this study had promising results, it had notable limitations. First, the training data was extracted from a single hospital. The extracted data was retrospective and collected in real life, which implies it contained biases, as the data reflected different practices and institutional adaptations of patient management protocols. Data extraction combined data from different software (DxCare and Clinysys GLIMS), which implied a potential loss of information, depending on the parameterization and coding of the various items of information entered, and data entry errors could have occurred. These errors and biases could have found their way into the final predictive model and been perpetuated. A critical analysis of the alerts issued by the algorithm is therefore essential. This is why a first processing of the data before model training was necessary to analyze and prepare the data for the training. Moreover, external validation using a similar dataset from other hospitals will help counter these biases and validate the models before their generalization.

Further studies must be conducted on the developed algorithms to test the models in real life and evaluate the clinical relevance of these approaches. PIs detected thanks to the algorithm will be compared to PIs detected through conventional medication review methods. Clinical and organizational impacts will also be studied to estimate the benefits of the use of ML in medication review activities. Finally, the deployment strategy for use of the model in a CDS system will need to be anticipated to avoid technical or organizational hurdles. The quality, interoperability, and seamless flow of data between software must be ensured, and the pharmacists' workflow process must evolve.

Conclusion

This study evaluated the performance of several ML models trained on a dataset (and subsamples) extracted from a single hospital to detect inappropriate prescription lines. The LGBM model was demonstrated to have the best overall performance, achieving higher accuracy, precision, specificity, F1-score, and AUC-ROC values when compared to the RF model. While further studies are needed to confirm these findings (by validating the model LGBM in other hospital settings to assess its generalizability and conducting prospective studies to evaluate the model's effectiveness in real-time clinical practice), the model holds

significant potential to advance hospital clinical pharmacy and improve patient care through optimized prescription management.

Acknowledgments

The authors thank T. Fabacher, MD, and J. Muller, MD, for extracting the data from the hospital information system.

Data availabilityThe datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Additional information

The authors provided the following CREdiT statement regarding their contributions to this work: Conceptualization: E.J., G.A., L.D.M., A.A., M.B., B.G., E.A.S., B.M.; Methodology: E.J.; Software: E.J.; Validation: L.D.M., M.B., B.G., E.A.S., B.M.; Formal analysis: E.J.; Investigation: E.J.; Resources: L.D.M., M.B., B.G., E.A.S., B.M.; Data curation: E.J.; Writing – Original draft: E.J., G.A., B.M.; Writing – Review & Editing: E.J., G.A., L.D.M., A.A., M.B., B.G., E.A.S., B.M.; Visualization: E.J., G.A.; Supervision: L.D.M., M.B., B.G., E.A.S., B.M.; Project administration: L.D.M., E.A.S., B.M.; Funding acquisition: L.D.M., E.A.S., B.M.

Disclosures

The authors have declared no potential conflicts of interest.

References

1. National Coordinating Council for Medication Error Prevention. Medication error definition. Accessed February 19, 2025. https://www.nccmerp.org/about-medication-errors
2. World Health Organization. Medication Without Harm. Accessed February 19, 2025. https://www.who.int/initiatives/medication-without-harm
3. World Health Organization. World Patient Safety Day 2022. Accessed February 19, 2025. https://www.who.int/campaigns/world-patient-safety-day/2022
4. Griese-Mammen N, Hersberger KE, Messerli M, et al. PCNE definition of medication review: reaching agreement. *Int J Clin Pharm*. 2018;40(5):1199-1208. doi:10.1007/s11096-018-0696-7
5. Bright JM, Tenni PC. The Clinical Services Documentation (CSD) System for documenting clinical pharmacists' services. *Aust J Hosp Pharm*. 2000;30(1):10-15. doi:10.1002/jppr200030110
6. Farhat A, Abou-Karroum R, Panchaud A, Csajka C, Al-Hajje A. Impact of pharmaceutical interventions in hospitalized patients: a comparative study between clinical pharmacists and an explicit criteria-based tool. *Curr Ther Res Clin Exp*. 2021;95:100650. doi:10.1016/j.curtheres.2021.100650
7. Novais T, Maldonado F, Grail M, Krolak-Salmon P, Mouchoux C. Clinical, economic, and organizational impact of pharmacists' interventions in a cognitive-behavioral unit in France. *Int J Clin Pharm*. 2021;43(3):613-620. doi:10.1007/s11096-020-01172-4
8. Zecchini C, Vo TH, Chanoine S, et al. Clinical, economic and organizational impact of pharmacist interventions on injectable antineoplastic prescriptions: a prospective observational study. *BMC Health Serv Res*. 2020;20(1):113. doi:10.1186/s12913-020-4963-7
9. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3(1):1-10. doi:10.1038/s41746-020-0221-y
10. Ranchon F, Chanoine S, Lambert-Lacroix S, Bosson J-L, Moreau-Gaudry A, Bedouch P. Development of artificial intelligence powered apps and tools for clinical pharmacy services: A systematic review. *Int J Med Inform*. 2022;172(104983). doi:10.1016/j.ijmedinf.2022.104983
11. Johns E, Alkanj A, Beck M, et al. Using machine learning or deep learning models in a hospital setting to detect inappropriate prescriptions: a systematic review. *Eur J Hosp Pharm Sci Pract*. Published online November 24, 2023. doi:10.1136/ejhpharm-2023-003857
12. Alkanj A, Godet J, Johns E, Gourieux B, Michel B. Deep learning application to automated classification of recommendations made by hospital pharmacists during medication prescription review. *Am J Health-Syst Pharm*. 2024;81(11):e296-e303. doi:10.1093/ajhp/zxae011
13. Hu Q, Wu B, Wu J, Xu T. Predicting adverse drug events in older inpatients: a machine learning study. *Int J Clin Pharm*. 2022;44(6):1304-1311. doi:10.1007/s11096-022-01468-7
14. Van Laere S, Muylle KM, Dupont AG, Cornu P. Machine learning techniques outperform conventional statistical methods in the prediction of high risk QTc prolongation related to a drug-drug interaction. *J Med Syst*. 2022;46(12):100. doi:10.1007/s10916-022-01890-4
15. Corny J, Rajkumar A, Martin O. et al. A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error. *J Am Med Inform Assoc*. 2020;27(11):1688–1694. doi:10.1093/jamia/ocaa154
16. Joseph VR. Optimal ratio for data splitting. *Stat Anal Data Min ASA Data Sci J*. 2022;15(4):531-538. doi:10.1002/sam.11583
17. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. *Data Min Knowl Discov*. 2014;28(1):92-122. doi:10.1007/s10618-012-0295-5
18. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
19. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451
20. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed February 19, 2025. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
21. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;12:5979. doi:10.1038/s41598-022-09954-8
22. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC Plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. doi:10.2307/2531595
24. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
25. Videau M, Charpiat B, Vermorel C, et al. Characteristics of pharmacist's interventions triggered by prescribing errors related to computerised physician order entry in French hospitals: a cross-sectional observational study. *BMJ Open*. 2021;11(10):e045778. doi:10.1136/bmjopen-2020-045778

26. Hamilton HJ, Gallagher PF, O'Mahony D. Inappropriate prescribing and adverse drug events in older people. *BMC Geriatr*. 2009;9:5. doi:10.1186/1471-2318-9-5
27. Soon HC, Geppetti P, Lupi C, Kho BP. Table 31.2. High-risk medication list. Published December 15, 2020. Accessed May 26, 2024. https://www.ncbi.nlm.nih.gov/books/NBK585602/table/ch31.Tab2/
28. Cornuault L, Mouchel V, Phan Thi TT, Beaussier H, Bézie Y, Corny J. Identification of variables influencing pharmaceutical interventions to improve medication review efficiency. *Int J Clin Pharm*. 2018;40(5):1175-1179. doi:10.1007/s11096-018-0668-y
29. Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. *Mach Learn*. 2021;110(3):451-456. doi:10.1007/s10994-021-05964-1
30. Balestra M, Chen J, Iturrate E, Aphinyanaphongs Y, Nov O. Predicting inpatient pharmacy order interventions using provider action data. *JAMIA Open*. 2021;4(3):ooab083. doi:10.1093/jamiaopen/ooab083
31. Imai S, Takekuma Y, Kashiwagi H, et al. Validation of the usefulness of artificial neural networks for risk prediction of adverse drug reactions used for individual patients in clinical practice. *PloS One*. 2020;15(7):e0236789. doi:10.1371/journal.pone.0236789
32. Hu Q, Wu B, Wu J, Xu T. Predicting adverse drug events in older inpatients: a machine learning study. *Int J Clin Pharm*. 2022;44(6):1304-1311. doi:10.1007/s11096-022-01468-7
33. Wongyikul P, Thongyot N, Tantrakoolcharoen P, Seephueng P, Khumrin P. High alert drugs screening using gradient boosting classifier. *Sci Rep*. 2021;11(1):20132. doi:10.1038/s41598-021-99505-4
34. Yalçın N, Kaşıkcı M, Çelik HT, et al. Development and validation of a machine learning-based detection system to improve precision screening for medication errors in the neonatal intensive care unit. *Front Pharmacol*. 2023;14:1151560. doi:10.3389/fphar.2023.1151560
35. Ben Othman S, Decaudin B, Odou P, Rousselière C, Cousein E, Hammadi S. Pharmaceutical decision support system using machine learning to analyze and limit drug-related problems in hospitals. *Stud Health Technol Inform*. 2024;310:1593-1597. doi:10.3233/SHTI231332
36. Alzubi J, Nayyar A, Kumar A. Machine learning from theory to algorithms: an overview. *J Phys Conf Ser*. 2018;1142(1):012012. doi:10.1088/1742-6596/1142/1/012012