Journal of Advanced Health Informatics Research (JAHIR)
Vol. 3, No. 1, April 2025, pp. 9-18
ISSN: 2985-6124, DOI: 10.59247/jahir.v3i1.341

Hybrid Ensemble Learning for Classifying Prescription vs. Over-the-Counter Medicines on Large-Scale Categorical and Textual Data

Reina Melani [1], Dina Febrina [2]
[1,2] *Department of Pharmacy, Harapan Bangsa University, Indonesia*

ARTICLE INFO

Article history:

Received February 28, 2025
Revised March 25, 2025
Published August 28, 2025

Keywords:

Prescription (Rx) drug;
Over-the-Counter (OTC) drug;
Interpretable CART;
TF-IDF;
LightGBM ensemble.

ABSTRACT

The classification of drugs into Prescription (Rx) and Over-the-Counter (OTC) categories is an important aspect of pharmaceutical governance because it has a direct impact on patient safety, drug access, and regulatory compliance. However, large-scale pharmaceutical data often consists of heterogeneous categorical variables and short texts, such as product names or indications, which poses challenges in the form of duplication, inconsistencies, and potential class imbalances. This condition demands a modeling approach that is not only accurate, but also lightweight and explainable. This study proposes a hybrid ensemble model that combines three algorithms, namely CART, Random Forest, and LightGBM, through a weighted soft-voting mechanism. This approach combines decision tree transparency with the reliability of modern boosting techniques. The main contribution of this study is to show that a low-complexity domain-based pipeline can produce accurate, efficient, and easily auditable Rx and OTC classifications for both clinical and regulatory needs. The pre-processing pipeline includes TF-IDF for short text, One-Hot Encoding for categorical features, as well as simple dosage variables. All features were combined into a solid matrix, then trained using weighted ensembles [1,1,8]. Evaluations include Accuracy, Precision, Recall, F1-score, ROC-AUC, Brier score, confusion matrix, and ROC curve. Test results on a dataset of 50,000 balanced samples showed consistent in-sample performance: Accuracy = 0.742; Accuracy = 0.742; Recall = 0.742; F1 = 0.742; ROC-AUC = 0.819; then Brier score = 0.214. The model is able to stably distinguish classes with a balance between False Positive and False Negative errors. In conclusion, this lightweight ensemble is able to present competitive prediction performance as well as interpretation, so that it has the potential to be applied to pharmacovigilance and drug classification. Further studies suggest adding cross-validation, probability calibration, as well as robustness tests to data outside the distribution to strengthen the reliability of the model.

Corresponding Author:

Reina Melani, Universitas Harapan Bangsa, Jl. Raden Patah No.100, Kec. Kembaran, Banyumas, Indonesia
Email: reinamelani@uhb.ac.id

1. INTRODUCTION

The classification of drugs into Prescription (Rx) and Over-the-Counter (OTC) categories is a pillar of pharmaceutical governance because it has a direct impact on patient safety, drug access, and regulatory compliance. Rx drugs can only be obtained through a doctor's prescription because their use requires diagnosis, clinical monitoring, or has a risk profile that demands the supervision of a healthcare professional [1]. OTC drugs can be purchased without a prescription for self-medication for mild conditions and are considered safe and effective when used according to the label [2]. Nonetheless, the risk of side effects, potential abuse, and variations in standards across jurisdictions demand strict governance [3]. Pharmaceutical data are generally categorical and short text at scale, posing particular challenges for modeling and evaluation.

In the context of implementation, pharmaceutical data is generally categorical and in the form of short text, e.g. drug names that often vary due to differences in trademarks and generic names, forms of preparations such as tablets, capsules, syrups, injections that can appear in different combinations, dosage or strength of drugs written in non-uniform formats, manufacturers with very diverse quantities and different production scales, as well as indications of use that tend to be brief but have the potential to overlap between categories [4]. These characteristics produce large amounts of data with a high degree of heterogeneity, often accompanied by duplication or inconsistencies in writing [5]. This complexity poses particular challenges in modeling, as the analysis method must be able to overcome variations in category representation, minimize the impact of distribution imbalances between classes, and maintain the consistency and reliability of prediction results even though the data is massive and not fully structured [6] .

This study used a lightweight ensemble that combined three algorithms, namely CART (Decision Tree), Random Forest, and LightGBM, through a weighted *soft-voting* mechanism. This approach was chosen to balance between predictive performance and model clarity [7]. CART provides a transparent and easily auditable decision path [8] , Random Forest and LightGBM contribute to consistently improving accuracy on large-scale text-category datasets [9]. Feature representations follow baseline design. For short texts, such as Name and Indication, TF-IDF is used, which is capable of highlighting important words in the context of prediction [10]. The main categorical features, including Category, Dosage Form, and Manufacturer, are encoded using One-Hot Encoding (OHE), so that each category is represented separately [11]. This study added simple dosage features, such as logarithmic transformation of dose and mild interactions between dose form and dose units, to enrich information related to drug use.

The evaluation of the model is carried out thoroughly. In addition to accuracy, the metrics used include Precision, Recall, F1-score, ROC-AUC, and Brier score, along with the Confusion Matrix and ROC curve as the main output. The default decision threshold of 0.5 is applied to all threshold-based classification metrics. The model does not use additional probability calibration, so the Brier score is used as a simple indicator to assess probability calibration. To prevent *information leakage* and maintain computational efficiency, all transformations of TF-IDF, OHE, and dose and interaction engineering features are only applied to the training data before being used on validation or test data. The validation scheme follows a computationally efficient baseline design, without the use of nested cross-validation. Overall, this baseline strategy emphasizes the achievement of a model that is lightweight, accurate, and easy to audit. The combination of TF-IDF, OHE, and simple dose features combined through *soft-voting* in the form of CART/RF/LGBM results in a good balance between performance and clarity, while maintaining the simplicity of the model so that it is easy to reproduce and interpret in the context of Prescription (Rx) and Over-the-Counter (OTC) drug classifications.

2. METHODS

2.1. Study Design

This study assesses the extent to which the use of the combined light ensemble method of several simple models can be used to classify the types of Prescription and OTC drugs on data dominated by categories such as the name of the manufacturer or the form of the preparation and short text such as the name of the drug or the indication of its use [12] . The three models used are CART, Random Forest (RF), and LightGBM (LGB). All three are combined with "soft-voting" methods, where LGBs are given greater weight because they are considered to contribute the most. Each model has certain initial settings, such as the number of decision trees, depth, and data sharing rules that generally aim to balance accuracy and overfitting risk [13]. In the preprocessing stage, text data such as Name and Indication are converted into numbers through the TF-IDF

technique so that they can be processed by computer, while category data such as Category or Dosage Form is converted into numerical code with One-Hot Encoding (OHE). The dosage information from Strength is also broken down into several simple forms (dose_log, dose_unit, dose_bucket) to ensure that the scale and variations can still be properly analyzed.

Evaluation of the performance of the baseline model was carried out after training on all data. The main results are seen from how well the model differentiates drug classes, as well as the general metrics of classification, namely accuracy, precision, recall, and F1-score. Brier Score is used to assess the accuracy of prediction probability [14]. The results of the analysis are visualized in the form of a Confusion Matrix table that shows the number of true and false predictions and an ROC curve that illustrates the balance of sensitivity and specificity [15] . To maintain transparency and for the study to be repeated, all results such as metric values, graphs, and hybrid models are stored as artifacts.

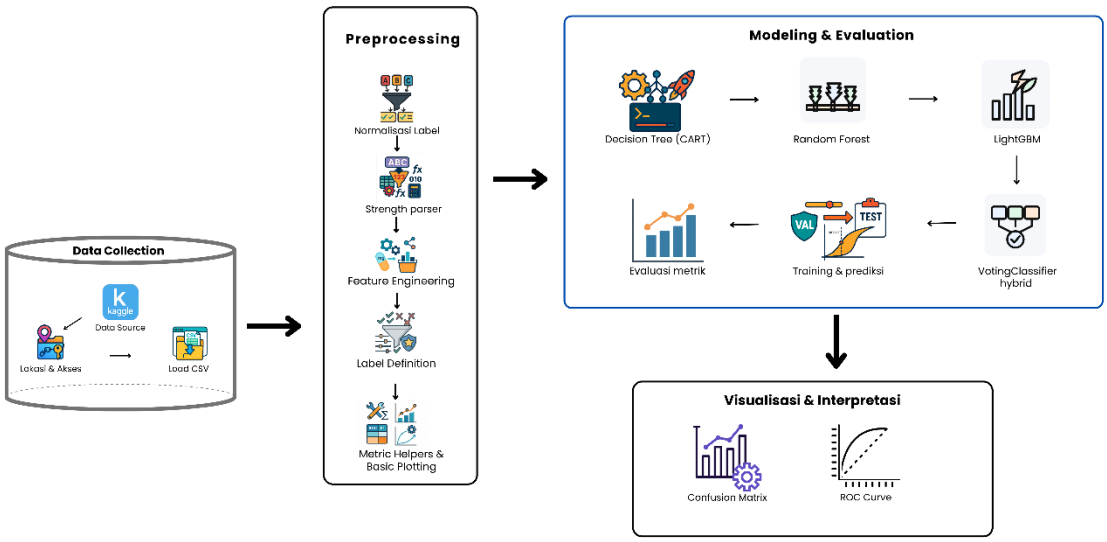

Fig 1. System Architecture

2.2 Dataset and Variables

The dataset used in this study is derived from Kaggle and contains pharmaceutical data in medium to large scales, which is mostly in the form of categories and text. The main variable predicted is Classification with two classes, namely Prescription (Rx) for prescription drugs, and Over-the-counter (OTC) for over-the-counter drugs. Since there are only two categories, this problem is formulated as a binary classification. The data includes important information such as product name, category of therapy, form or route of preparation, dosage or strength of the drug, manufacturer, as well as clinical indications. In accordance with the basic needs of the research, dose information has been prepared in the form of structured variables such as dose_log, dose_unit, and dose_bucket so that it is ready to be used in modeling. Given that data is diverse and sometimes repetitive, the next steps are focused on standardizing writing and encoding the categorical data consistently. All processes are carefully carried out to prevent information leakage between training and testing data, so that the evaluation results remain valid.

2.3 Data Preprocessing and Cleaning

The preprocessing stage is carried out to standardize categorical data and short text, with all steps designed to avoid information leakage between the training data and the test data (in-fold). The process begins with normalizing the text on attributes such as Name, Category, Dosage Form, Manufacturer, and Indication, which includes lowercase equalization, redundancy of excess space, and standardization of hyphens [16] [17] . In accordance with the baseline design, the dosage information that has been described before, namely dose_log, dose_unit, and dose_bucket is assumed to be available so that at this stage there is no further decomposition from the Strength column. For missing data, a simple strategy is used: category values or text are filled with

"Missing" tokens, while numerical values are filled in using the median of the training data in each fold. After that, the data is separated with a stratification scheme to keep the class distribution balanced, and if necessary, manufacturer-based separation can be carried out to test the robustness of the model.

According to the baseline, the text was processed using the TF-IDF method which calculates the importance of words based on the frequency of their occurrence. This process is implemented separately for Names and Indications with specific configurations maximum feature count of 3500, word combinations of 1–6 words, and common word exclusions in English. Key categorical features such as Category, Dosage Form, and Manufacturer are converted into numerical representations using One-Hot Encoding (OHE), which is a technique that assigns special marks to each category [18]. Additional domain-relevant interactions are created, Dosage Form × dose_unit combinations, which are also processed with OHE. All pre-processing outputs, both OHE results, numerical dose variables, and TF-IDF matrices, are then combined into one final feature matrix. This matrix is designed to remain in accordance with the needs of the baseline model (CART, RF, LGB) and can be used directly at the training and evaluation stages.

2.4 Domain-Aware Feature Engineering

The feature engineering stage is focused on important clinical information without adding unnecessary complexity, according to the baseline design. Dose information is considered available in a structured form, especially dose_log, dose_unit, and dose_bucket so that there is no need to re-decompose the Strength column. To capture the relationship between the form/route of the preparation and the dosage unit, a simple combination (Dosage Form × dose_unit) was used which was then converted into a numerical format. The main categorical features (Category, Dosage Form, Manufacturer) are changed using OHE, which is a technique that codes each category specifically so that it can be processed by the model [19] . Short texts such as Name and Indication are processed by the TF-IDF method which measures how important a word is based on its frequency and is processed separately [20] . All of these pre-processing steps are carefully carried out to prevent information leakage between the training and test data, and then combined into a single feature matrix that is ready to be used by the CART, RF, and LGB models according to the baseline.

As an additional option at the tuning stage, combinations between other relevant categorical variables such as Category × Dosage Form or adding dose_bucket to an existing interaction can be tried gradually. The impact of these additions was evaluated by looking at changes in model performance measures, such as class differentiation ability and predictive probability accuracy (Brier score), without changing the main configuration of the TF-IDF baseline for text, OHE for categories, as well as the CART + RF + LGB ensemble with weighted soft-voting.

2.5 Categorical Feature Encoding

In the baseline design, categorical data is changed using OHE, which is a technique that provides binary code for each category so that the model can be processed. This process is applied to the Category, Dosage Form, and Manufacturer attributes, with settings so that new, unknown categories don't cause errors. For data in the form of short text, namely Name and Indication, the TF-IDF method is used, which assesses how important a word is based on its frequency in the data. These two columns are processed separately with specific configurations (maximum 3,500 features, word combinations of up to six words in a row, and ignore English stop words). The results of OHE, TF-IDF, and dose-related numerical features are then combined into a dense data matrix and ready to be used by CART, RF, and LGB models according to the baseline design.

As an additional option for experiments, techniques such as rare-bucketing combining very rarely appearing categories into a single group [21] , or Target Encoding (TE) converting categories into numerical values based on relationships to targets can be considered for features where the number of categories is very large [22]. If used, this technique must be applied carefully to the training data of each fold (in-fold) so that there is no information leakage between the training and test data. The main configuration of the study remains using OHE for categorical data and TF-IDF for short text.

2.6 Classification Models

The classification model used in this study is in the form of light ensembles, which are a combination of several models to complement each other's strengths. The three models combined are a simple CART decision tree that is easy to understand [23] , a Random Forest (RF) combination of many decision trees to improve generalization [24] , and LightGBM (LGB) a gradient-based model that is faster and more efficient [25]. All

three models are combined with the soft-voting method, where the final prediction is determined based on the average probability of each model. The initial weight given is [1, 1, 8], so LGB has a greater influence because it generally provides the most accurate results with light computing. All models run using the same data (dense features of OHE and TF-IDF), so no additional conversion is required.

The basic settings of each model follow a baseline configuration of CART limited tree depth to keep it simple and balanced [26] , RF uses multiple decision trees with settings to avoid bias [27] , while LGB is set with a similar number of trees, limited depth, and a certain learning speed to maintain accuracy as well as efficiency [28] . The baseline training process is carried out using all available data, while voting weight adjustments are only made if needed at the advanced stage (tuning). Parameter adjustments (hyperparameter tuning) can also be done gradually, focusing on key factors such as tree depth, minimum amount of data per node, number of trees, and learning rate. In this baseline, no additional calibration is carried out on the prediction probability to remain efficient; Probability values are taken directly from the model.

2.7 Evaluation Protocols

Evaluation at baseline was carried out after all models (CART, RF, LGB) and the data preparation process was trained using all available data. The data is processed into several types of features, including the text in the Name and Indication is converted into numbers using the TF-IDF method, while category data such as Category, Dosage Form, and Manufacturer are changed with OHE so that it can be read by machines. Additional information such as drug dosage (dose_log, dose_unit, dose_bucket) is also used as a feature, including a simple combination of the drug preparation form and the dosage unit. After all the features were combined, the combined model (ensemble soft-voting) with the initial weights [1,1,8] was fully trained, and then predictions were made on the same data to calculate the performance size.

The main performance reported was the ROC AUC, as it was able to illustrate the overall accuracy of the model. In addition, other standard measures such as Accuracy, Precision, Recall, and F1 are also used, as well as an additional measure to assess the reliability of the predictive probability (Brier score). The results of the evaluation are visualized in the form of a Confusion Matrix (heatmap) and a ROC curve. Classification decisions are determined with a standard threshold of 0.5, as per common practice at baseline. The saved final result includes metric values, two main visualizations, and a reusable trained model (hybrid_model) object. Additional evaluation methods such as cross-validation, leave-manufacturer-out testing, or bootstrap confidence interval calculation are not included in the baseline, but can be added at a later stage of research if needed.

3. RESULTS AND DISCUSSION

3.1. Dataset Profile and Preprocessing Outcomes

The dataset consisted of 50,000 entries with a balanced label distribution between Prescription (Rx) and Over-the-Counter (OTC), so that the estimation of discriminatory performance was not affected by class inequality. This balance is important because it makes the results of the model evaluation unbiased to one of the classes. The data has many variations, especially in the information of the manufacturer (Manufacturer), category of therapy (Category), and dosage form (Dosage form). Meanwhile, the information in the product name and usage indication columns (Indication) is often written in different styles, such as differences in uppercase letters or the use of hyphens. Therefore, data cleaning is carried out by equalizing the text format to make it neater and more consistent. Data on drug doses (dose_log, dose_unit, dose_bucket) are available from the beginning, so there is no longer a need to process them from the Strength column.

According to the baseline, the text in the Name and Indication columns is converted into numbers using a special method that assesses how important a particular word is in the entire dataset. The main category information such as Category, Dosage Form, and Manufacturer is also converted to numerical form so that it can be processed by the computer. In addition, a simple combination of the dosage form and dosage units is made as an additional feature. All data processing results from text, categories, and doses are then combined into a single unit so that they are ready to be used to train classification models, both single models (CART, Random Forest, LightGBM) and combined models (soft-voting ensemble) according to the research design.

3.2 Training Summary

In the initial stage, the training was carried out directly on all data (**in-sample**) with a matrix of features that had been combined in solid form. This matrix consists of: category data (Category, Dosage Form, Manufacturer) modified with One-Hot Encoding (OHE), short text (Name and Indication) processed using TF-IDF (maximum 3500 important words, combination of 1–6 words, with English stop words), numerical dosage feature (dose_log), and simple interaction between Dosage Form and dose_unit which is also encoded with OHE. The three main CART models, Random Forest (RF), and LightGBM (LGB) were then combined with a soft-voting method using weights [1, 1, 8], where LGBs were given greater weight because they were considered to contribute the most. Baseline evaluation was carried out on all data with a standard decision threshold of 0.5.

3.3 In-Sample Performance

After all the pre-processing stages of the data were applied, a hybrid model with the weighted soft-voting method [1, 1, 8] was tested using all available data. This pre-processing process includes encoding category data such as Category, Dosage Form, and Manufacturer into numerical format, processing short texts such as Name and Indication using the TF-IDF method (with a combination of 1–6 words), as well as adding simple dose information (dose_log) and interaction between the dosage form and the dosage unit. Evaluation was carried out with a standard decision threshold of 0.5. Reported performance measures include Accuracy, Precision, Recall, F1-score, ROC-AUC, and Brier score. Since this test was performed on the same data as the training data (in-sample), the results obtained mainly illustrate the model's ability to adapt to the data, so keep in mind the risk of overfitting.

In a dataset of 50,000 samples, the hybrid model showed the following results: Accuracy = 0.742; Accuracy = 0.742; Recall = 0.742; F1 = 0.742; ROC-AUC = 0.819; then Brier score = 0.214. These results indicate that the model is able to differentiate classes well (indicated by the high AUC value) and maintain a balance between precision and recall in training data. However, these findings are still indicative of fit, not a guarantee of the same performance when faced with new data. In the next stage, Macro-F1 can be added as a complementary measure to provide a more comprehensive picture.

3.4 Confusion Matrix

Fig 2. shows the confusion matrix of the hybrid model soft-voting, w = [1,1,8] at the decision threshold 0.5 with class labels 0 and 1. Out of a total of 50,000 data, the correct predictions were recorded on the diagonal 18,532 for class 0 and 18,525 for class 1. The distribution of misclassification was almost evenly distributed, namely False Positive (FP) as many as 6,483 (class 0 predicted as 1) and False Negative (FN) as many as 6,460 (class 1 predicted as 0).

This pattern resulted in an equivalent recall value in both classes, which was around 0.741 for class 0 (18,532/25,015) and 0.741 for class 1 (18,525/24,985). Likewise, the precision value per class is relatively similar, which is around 0.742 for class 0 and 0.741 for class 1. These results are in line with the summary of the previously presented in-sample metrics. From a practical point of view, the balance of the number of FP and FN shows that at the threshold of 0.5 the model does not show bias towards either class. However, in real application, FN errors may be considered more risky for example, misclassifying drugs that should be closely monitored. In that context, decision threshold adjustments or weight additions to certain classes can be considered at a later stage of development. These adjustments are not included in the scope of the in-sample baseline evaluation reported here.

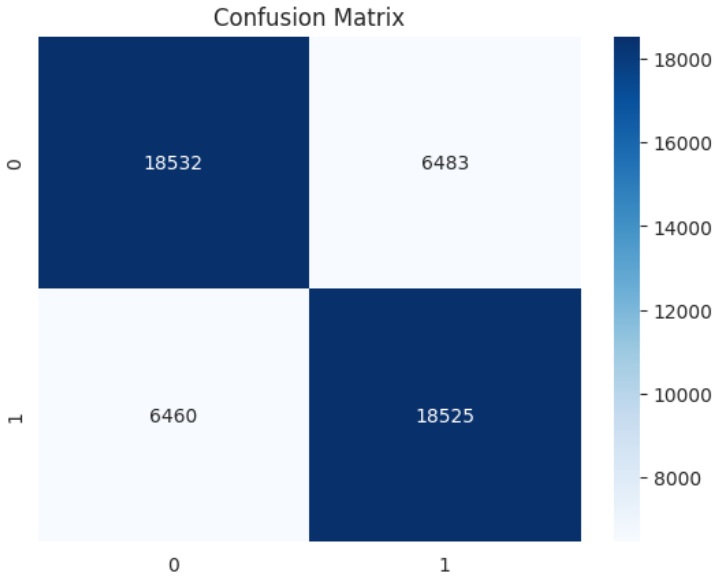

Fig 2. Confusion Matrix

3.5 ROC Analysis

At the evaluation stage, the combined model (soft-voting ensemble) with an initial weight (1,1,8) showed good performance in distinguishing between prescription drugs (Rx) and over-the-counter drugs (OTC). The ROC AUC value achieved is 0f.819, indicating a fairly high ability to distinguish the two classes. ROC curve graph Figure 3. It looks clearly above a random line (dotted line), which means the model can separate classes fairly stably across various decision thresholds.

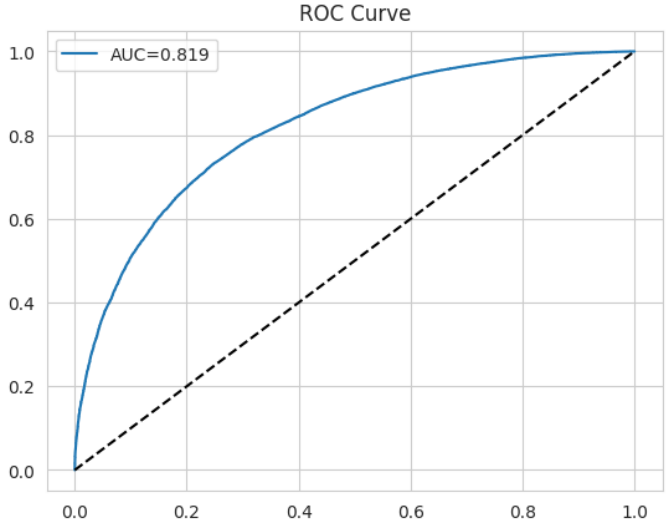

Fig 3. ROC Curve Graph on Hybrid Model

The ROC curve graph in Fig. 3. shows that the combined model's performance is well above the random line, so the resulting classification decisions are relatively consistent across multiple thresholds. In simple terms, this suggests that the combination of information from product names and indications for use (processed with TF-IDF), as well as categorical features such as drug category, dosage form, manufacturer (processed with One-Hot Encoding), and available dosage data, is strong enough to distinguish between prescription and over-the-counter drugs in this scenario. According to the scope of baseline research, the assessment is focused on the ROC AUC.

4. CONCLUSION

A lightweight and interpretable approach to the classification of Prescription (Rx) versus Over-the-Counter (OTC) drugs on large datasets demonstrates competitive performance while being easy to audit. The pipeline used combines TF-IDF for Name and Indication columns, One-Hot Encoding for Category, Dosage Form, and Manufacturer, as well as additional simple dosing features. A soft-voting ensemble consisting of CART, Random Forest, and LightGBM with weights [1,1,8] was trained on 50,000 samples, resulting in stable in-sample performance: Accuracy 0.742; Precision 0.742; Recall 0.742; F1 0.742; ROC-AUC 0.819; and Brier's score of 0.214. These results show a balance between prediction accuracy, computational efficiency, and model interpretability.

However, these results need to be interpreted carefully because the evaluation is carried out in-sample, so it tends to provide a more optimistic picture of performance. The risk of overfitting still exists, while the evaluation of probability calibration such as ECE or reliability curve or Decision Curve Analysis (DCA) has not been carried out. The model's resilience to out-of-distribution scenarios, such as leave-manufacturer-out or time-split tests, has not been tested. The assumption that structured dosing features (dose_log, dose_unit, dose_bucket) are always available also needs to be verified, as variations in more complex dosing formats could affect the results.

For further development, it is recommended to implement stricter validation, such as nested cross-validation with bootstrap-based trust interval reporting. The determination of decision thresholds should be adjusted to practical or policy needs. Efficient probability calibration using the Platt or Isotonic method can be enriched by reporting additional metrics such as ECE, and reliability curves. Resistance tests to distribution shifts and feature ablation analysis can also help balance model accuracy and complexity.

In the context of implementation in resource-constrained environments, lighter single-model usage options may be prioritized, such as trimmed or calibrated LightGBM, or simplified CART with post-process rules. This step should be accompanied by improvements to label normalization and rare category handling to keep the pipeline stable and replicable. With this strategy, the model has the potential to evolve from a mere proof of concept to a more reliable, scalable, and auditable solution, both in clinical and regulatory contexts.

REFERENCES

[1] A. Yasmeen *et al.*, "Suspected inappropriate use of prescription and non-prescription drugs among requesting customers: A Saudi community pharmacists' perspective," *Saudi Pharmaceutical Journal*, vol. 31, no. 7, pp. 1254–1264, Jul. 2023, doi: 10.1016/j.jsps.2023.05.009.

[2] A. Hatabu, Y.-S. Tian, H. Asano, K. Fukuzawa, and K. Ikeda, "A brief report of the status of self-medication with over-the-counter drugs: a pilot cross-sectional survey," *BMC Res Notes*, vol. 18, no. 1, p. 37, Jan. 2025, doi: 10.1186/s13104-025-07114-5.

[3] E. Toni, H. Ayatollahi, R. Abbaszadeh, and A. Fotuhi Siahpirani, "Machine Learning Techniques for Predicting Drug-Related Side Effects: A Scoping Review," *Pharmaceuticals*, vol. 17, no. 6, p. 795, Jun. 2024, doi: 10.3390/ph17060795.

[4] W. Guo, F. Dong, J. Liu, A. Aslam, T. A. Patterson, and H. Hong, "A refined set of RxNorm drug names for enhancing unstructured data analysis in drug safety surveillance," *Exp Biol Med*, vol. 250, May 2025, doi: 10.3389/ebm.2025.10374.

[5] S. Janiczak *et al.*, "An Evaluation of Duplicate Adverse Event Reports Characteristics in the Food and Drug Administration Adverse Event Reporting System," *Drug Saf*, vol. 48, no. 10, pp. 1119–1126, Oct. 2025, doi: 10.1007/s40264-025-01560-7.

[6] S. Dimitsaki, P. Natsiavas, and M.-C. Jaulent, "Applying AI to Structured Real-World Data for Pharmacovigilance Purposes: Scoping Review," *J Med Internet Res*, vol. 26, p. e57824, Dec. 2024, doi: 10.2196/57824.

[7] K. Cao-Van, T. C. Minh, L. G. Minh, T. T. B. Quyen, and H. M. Tan, "Soft-Voting Ensemble Model: An Efficient Learning Approach for Predictive Prostate Cancer Risk," *Vietnam Journal of Computer Science*, vol. 11, no. 04, pp. 531–552, Nov. 2024, doi: 10.1142/S2196888824500155.

[8] A. Argente-Garrido, C. Zuheros, M. V. Luzón, and F. Herrera, "An Interpretable Client Decision Tree Aggregation process for Federated Learning," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.02510

[9] Z. Wang, H. Ren, R. Lu, and L. Huang, "Stacking Based LightGBM-CatBoost-RandomForest Algorithm and Its Application in Big Data Modeling," in *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/DOCS55193.2022.9967714.

[10] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," *Array*, vol. 27, p. 100467, Sep. 2025, doi: 10.1016/j.array.2025.100467.

[11] Y. Zhang, L. He, Y. Zhang, P. Zhao, B. Zhang, and F. Cheng, "A comparative study of One-Hot,TF-IDF, and Word2Vec for Classifying Illegal Advertising Texts," in *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, New York, NY, USA: ACM, Dec. 2024, pp. 82–86. doi: 10.1145/3711542.3711586.

[12] B. Erdebilli and B. Devrim-İçtenbaş, "Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey," *Mathematics*, vol. 10, no. 14, p. 2466, Jul. 2022, doi: 10.3390/math10142466.

[13] E. Mahamud, M. Assaduzzaman, J. Islam, N. Fahad, M. J. Hossen, and T. T. Ramanathan, "Enhancing Alzheimer's disease detection: An explainable machine learning approach with ensemble techniques," *Intell Based Med*, vol. 11, p. 100240, 2025, doi: 10.1016/j.ibmed.2025.100240.

[14] W. Yang, J. Jiang, E. M. Schnellinger, S. E. Kimmel, and W. Guo, "Modified Brier score for evaluating prediction accuracy for binary outcomes," *Stat Methods Med Res*, vol. 31, no. 12, pp. 2287–2296, Dec. 2022, doi: 10.1177/09622802221122391.

[15] K. J. Sowmiya Narayanan and A. Manimaran, "Using Decision Risk and Decision Accuracy Metrics for Decision Making for Remote Sensing and GIS Applications," 2024, pp. 125–136. doi: 10.1007/978-981-99-6229-7_11.

[16] R. Jevsejev, D. Mažeika, and M. Bereiša, "An Approach for Building IT Support Dataset for Machine Learning Models," in *2025 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/eStream66938.2025.11016852.

[17] I. Hasan and M. Tausif, "Designing an Interpretable and Efficient AutoML Pipeline for Enhanced Data Analytics," in *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, IEEE, Jun. 2025, pp. 911–916. doi: 10.1109/ICSSAS66150.2025.11081354.

[18] A. M, N. Savarimuthu, and S. M. S. Bhanu, "WoEEE: a hybrid approach for enhancement of categorical data transformation," *Int J Data Sci Anal*, vol. 20, no. 7, pp. 6635–6663, Nov. 2025, doi: 10.1007/s41060-025-00845-5.

[19] T. Al-Shehari and R. A. Alsowail, "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques," *Entropy*, vol. 23, no. 10, p. 1258, Sep. 2021, doi: 10.3390/e23101258.

[20] E. Aljohani, "Enhancing Arabic Text Classification with a Hybrid Word Embedding Method," in *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, Dec. 2023, pp. 696–701. doi: 10.1109/DeSE60595.2023.10468772.

[21] D. Zhou and J. He, "Rare Category Analysis for Complex Data: A Review," *ACM Comput Surv*, vol. 56, no. 5, pp. 1–35, May 2024, doi: 10.1145/3626520.

[22] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Comput Stat*, vol. 37, no. 5, pp. 2671–2692, Nov. 2022, doi: 10.1007/s00180-022-01207-6.

[23] A. Abedinia and V. Seydi, "Building semi-supervised decision trees with semi-cart algorithm," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 10, pp. 4493–4510, Oct. 2024, doi: 10.1007/s13042-024-02161-z.

[24] N. E. I. Karabadji, A. Amara Korba, A. Assi, H. Seridi, S. Aridhi, and W. Dhifli, "Accuracy and diversity-aware multi-objective approach for random forest construction," *Expert Syst Appl*, vol. 225, p. 120138, Sep. 2023, doi: 10.1016/j.eswa.2023.120138.

[25] J. Huang and W. Chen, "A Study on Category Classification Based on LightGBM for Signal Feature Extraction and K-Means Clustering," in *2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, IEEE, Jul. 2023, pp. 858–862. doi: 10.1109/ICPICS58376.2023.10235522.

[26] R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales, "Optimal randomized classification trees," *Comput Oper Res*, vol. 132, p. 105281, Aug. 2021, doi: 10.1016/j.cor.2021.105281.

[27] T. T. Tran, N. Q. Phan, and H. X. Huynh, "Random Forest Model Parameters Optimization," 2025, pp. 237–247. doi: 10.1007/978-981-97-9616-8_19.

[28] S. Li, N. Jin, A. Dogani, Y. Yang, M. Zhang, and X. Gu, "Enhancing LightGBM for Industrial Fault Warning: An Innovative Hybrid Algorithm," *Processes*, vol. 12, no. 1, p. 221, Jan. 2024, doi: 10.3390/pr12010221.