Indonesian Journal of Data and Science

Volume 7 Issue1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i1.397

Research Article

Drug Recommendation Using Multilabel Classification with Decision Tree Based on Patient Complaints and Diagnoses

Muh Aristsyah Malik [1,*] ; Harlinda [2] ; Herdianti Darwis [3]
[1] Universitas Muslim Indonesia, Makassar, 90231, Indonesia, muharistsyahmalik@umi.ac.id
[2] Universitas Muslim Indonesia, Makassar, 90231, Indonesia, harlinda@umi.ac.id
[3] Universitas Muslim Indonesia, Makassar, 90231, Indonesia, herdianti.darwis@umi.ac.id

Correspondence should be addressed to Muh Aristsyah Malik; muharistsyahmalik@umi.ac.id

Received 13 February 2026; Accepted 30 March 2026; Published 30 March 2026

Abstract:
This study develops a drug recommendation system using multilabel classification with the Decision Tree algorithm based on patient complaint and diagnosis data from electronic medical records. The dataset consists of patient visit records from a community health center in Pangkajene and Kepulauan Regency and is transformed using multi-hot encoding. Model performance is evaluated under three dataset scenarios (N=500, N=800, and N=1000) using multilabel metrics, including Micro-F1, Samples-F1, Hamming Loss, Jaccard Similarity, Hit@5, Precision@K, and Recall@K. The best Decision Tree model achieved a Micro-F1 score of 0.292, Samples-F1 of 0.281, and Hit@5 of 0.690 on the N=1000 dataset scenario. Bootstrap validation with 1000 iterations indicates relatively stable performance, with narrow confidence intervals across evaluation metrics. These results show that the multilabel Decision Tree model is capable of capturing relationships between patient complaints, diagnoses, and drug therapies while maintaining an interpretable decision structure.

Keywords: Multilabel Classification; Decision Tree; Drug Recommendation System; Electronic Medical Records; Clinical Decision Support.

Dataset link: BPJS Drug Recommendation Dataset at the Bonto Perak Community Health Center, Pangkep Regency, South Sulawesi.

1. Introduction

The rapid advancement of information technology has encouraged extensive data utilization across various sectors, including healthcare [1]. One significant application is the use of electronic medical record (EMR) data to support clinical decision-making processes [2]. In healthcare practice, particularly in primary healthcare facilities such as community health centers, medical personnel frequently encounter patients presenting with multiple complaints and diagnoses simultaneously [3]. As a consequence, the prescribed therapy often consists not of a single medication, but rather a combination of several drugs tailored to the patient's clinical condition [4].

Drug recommendation systems represent a promising solution to assist healthcare professionals in determining appropriate therapeutic interventions based on patient data [5], [6]. However, most existing approaches model drug recommendation problems as single-label classification tasks, where each patient case is associated with only one output label [7]. This assumption does not adequately reflect real-world clinical settings, where the relationship between complaints, diagnoses, and therapeutic drugs is inherently many-to-many [8]. Therefore, a multilabel classification approach is required to simultaneously predict multiple drug labels within a single modeling process [9].

Machine learning techniques have been widely applied to classification and recommendation problems in healthcare due to their ability to learn patterns from historical data [10], [11], [12], [13]. Among various algorithms, Decision Tree offers a significant advantage in terms of interpretability. This algorithm constructs a model in the form of a decision tree structure that is logically explainable and easily understood [14]. Interpretability is particularly critical in medical decision support systems, as generated recommendations must be transparent and traceable by healthcare professionals [15].

Despite its interpretability advantages, applying Decision Tree to multilabel classification problems presents several challenges, including high feature dimensionality resulting from multilabel representation, imbalanced distribution of drug therapy labels, and the risk of overfitting if model parameters are not properly controlled [16]. Furthermore, evaluating multilabel-based drug recommendation systems cannot rely solely on a single accuracy metric; instead, it requires specialized evaluation measures capable of comprehensively assessing predictive accuracy and recommendation relevance [17].

Based on these challenges, this study proposes a multilabel drug recommendation system using the Decision Tree algorithm by leveraging patient complaint and diagnosis data as input features [18]. The study evaluates model performance across different dataset size scenarios to analyze the impact of data scale on predictive performance. The evaluation is conducted using multiple multilabel metrics, including Micro-F1, Samples-F1, Hamming Loss, Jaccard Similarity, Hit@5, Precision@K, and Recall@K, and Subset Accuracy [19]. In addition, statistical validation is performed using the bootstrap method to assess the stability and reliability of evaluation results. It is expected that this research will contribute to the development of an accurate, stable, and interpretable drug recommendation system to support data-driven clinical decision-making [20].

2. Method

This study employed an experimental approach to develop a multilabel drug recommendation system using the Decision Tree algorithm [21]. The entire process was implemented in Python using the Scikit-learn library, including multilabel data preprocessing, feature transformation through multi-hot encoding, model construction and optimization, performance evaluation using multilabel metrics, and statistical validation via the bootstrap method [22].

Experiments were conducted under three dataset size scenarios: 500, 800, and 1000 patient records. For each scenario, the dataset was split into training and testing sets with an 80:20 ratio. Because the dataset was collected at the visit level and did not include explicit patient identifiers, each record was treated as an independent clinical encounter. The data therefore did not contain longitudinal linkages between visits that could introduce cross-sample dependency. Under this structure, a random 80:20 train–test split was considered appropriate [23], [24]. In addition to scenario-based experiments, a global Decision Tree model was trained using the full dataset to enable visualization of the tree structure and extraction of decision rules [25]. The overall research workflow is illustrated in **Figure 1**.

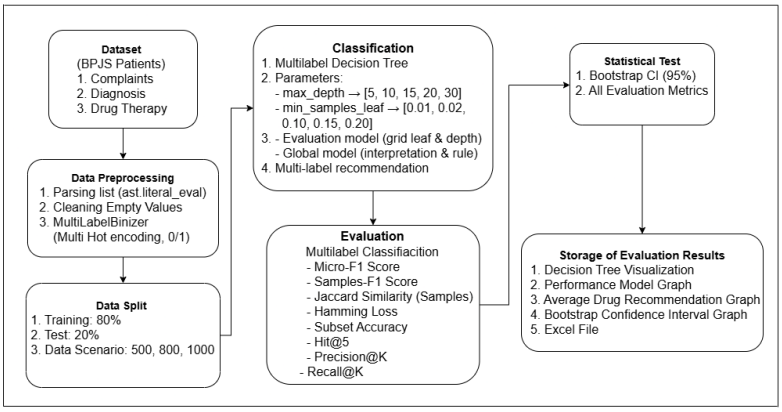

Figure 1. Research workflow of the multilabel classification drug recommendation system.

Dataset

The dataset used in this study was obtained from patient visit records at a community health center located in Pangkajene and Kepulauan Regency (Pangkep). The dataset consists of three primary attributes: patient complaints, patient diagnoses, and prescribed drug therapies [26]. Each attribute is represented in multilabel format, meaning that a single patient record may contain multiple complaints, multiple diagnoses, and multiple prescribed medications [27].

Data collection was conducted by tracing patient visit codes for specific dates through the health center information system. The data were accessed via the *Data Entry* > *Registration* menu, after which relevant clinical attributes were recorded and transferred into spreadsheet format [28]. This study focused exclusively on complaint, diagnosis, and drug therapy attributes. The dataset was stored in .xlsx format and processed using the pandas library in Python. To analyze the impact of dataset size on model performance, three dataset scenarios were defined: 500 (500 records), 800 (800 records), and 1000 (1000 records). **Table 1** shows several examples of patient visit records used in this study. The table illustrates how each record may contain multiple patient complaints, diagnoses, and prescribed drug therapies, reflecting the multilabel nature of the dataset used for model development.

Table 1. Dataset

No.	Patient Complaints	Patient Diagnoses	Prescribed Drug Therapies
1	['fever', 'common cold', 'headache']	["hyperlipidaemia"]	['vitamin B complex', 'simvastatin']
2	['cough', 'fever', 'common cold']	["fever"]	['vitamin C', 'cefadroxil', 'chlorpheniramine', 'paracetamol']
3	['joint pain']	["essential (primary) hypertension", "hyperlipidaemia"]	['allopurinol', 'diclofenac sodium', 'amlodipine', 'simvastatin']
4	['neck stiffness']	["non-insulin-dependent diabetes mellitus without complications", "hyperlipidaemia"]	['diclofenac sodium', 'glimepiride', 'cyanocobalamin (vitamin B12)', 'simvastatin']
5	['headache']	["vertigo of central origin"]	['paracetamol', 'betahistine', 'cyanocobalamin (vitamin B12)']

Data Preprocessing

The preprocessing stage was conducted to prepare the dataset for multilabel classification modeling. The dataset was first loaded from spreadsheet files using pandas and consisted of three main attributes: complaints, diagnoses, and drug therapies.

Each entry within these attributes was parsed using the ast.literal_eval function to convert textual list representations into valid Python list structures. During this process, extra whitespace was removed, and missing or invalid values were converted into empty lists to maintain structural consistency.

Subsequently, as presented in **Table 2**, the cleaned multilabel data were transformed into numerical representations using the MultiLabelBinarizer technique. Complaint and diagnosis attributes were converted into binary feature vectors using multi-hot encoding, serving as input features. Meanwhile, the drug therapy attribute was encoded as multilabel output targets. The encoding process resulted in 64 unique complaint features and 140 diagnosis features, forming a total of 204 binary input variables. For the output space, 69 distinct drug therapy labels were identified. The average label cardinality was 3.777, indicating that each patient received approximately four medications per visit, while the label density was 0.0547, reflecting a sparse multilabel output structure. Such sparsity increases the difficulty of exact multilabel prediction under strict evaluation metrics such as Subset Accuracy. This transformation produced a numerical representation suitable for training and evaluating the Decision Tree model [29].

Table 2. Data Preprocessing

Processing Stage	*Attribute*	*Original Data Example (Spreadsheet)*	*Processed Output*	*Description*
List Parsing	Complaints	['Fever', ' Cough ', 'Cold']	[Fever, Cough, Cold]	Text entries are converted into list format and extraneous whitespace is removed.
	Diagnosis	['ISPA']	[ISPA]	Single-item lists are preserved without modification.
	Drug Therapy	['Paracetamol', ' Amoxicillin ', 'CTM']	[Paracetamol, Amoxicillin, CTM]	Multiple items are separated and treated as multilabel outputs.
Missing Value Handling	Complaints	NaN	[]	Missing values are converted into empty lists.
	Diagnosis	' '	[]	Ensures structural consistency of the dataset.
MultiLabel Binarizer	Complaints	[Fever, Cough, Cold]	[1, 1, 1, 0, 0, …]	Multi-hot encoding is applied to complaint features.
	Diagnosis	[ISPA]	[0, 1, 0, 0, …]	Binary representation of diagnosis labels.
	Drug Therapy	[Paracetamol, Amoxicillin, CTM]	[1, 0, 1, 1, 0, …]	Multilabel output representation.

Classification

The classification stage in this study employed a Multilabel Decision Tree to model the relationship between patient complaints and diagnoses and the prescribed drug therapies. The model was developed using the DecisionTreeClassifier algorithm, which is capable of producing more than one output label for each patient record.

Model evaluation was conducted using several combinations of hyperparameters, specifically max_depth (5, 10, 15, 20, 30) and min_samples_leaf (0.01, 0.02, 0.10, 0.15, 0.20), with the objective of achieving a balance between model complexity and generalization capability [30]. Based on the evaluation results, the best-performing model configuration was obtained with max_depth = 15 and min_samples_leaf = 0.02.

Each parameter configuration was evaluated using multilabel classification metrics to determine the best-performing model. In addition, a global Decision Tree model was trained using the entire dataset for interpretability purposes, including visualization of the tree structure and extraction of decision rules. The final outcome of this stage is a multilabel drug recommendation system capable of generating more than one therapeutic recommendation simultaneously, based on combinations of patient complaints and diagnoses [31]. **Table 3** presents several examples of the classification results obtained under different dataset size scenarios and model parameter configurations. The table illustrates how the multilabel Decision Tree model generates drug therapy recommendations based on combinations of patient complaints and diagnoses.

Table 3. Multilabel Classification Based on Data Volume Scenarios and Model Parameters

Data	*Min samples leaf*	*Max depth*	*Patient Complaints*	*Patient Diagnoses*	*Recommended Drug Therapies*	*Number of Drugs*
500	0.10	30	Shortness of breath, Joint pain	Low back pain	Ibuprofen, Cotrimoxazole (sulfamethoxazole–trimethoprim combination)	5
800	0.01	20	Headache	Essential (primary) hypertension, hyperlipidaemia	Betahistine, Simvastatin	2
1000	0.02	15	Eye swelling	Hordeolum and other deep	Attapulgite, Oral rehydration salts (sodium chloride-based	6

Data	*Min samples leaf*	*Max depth*	*Patient Complaints*	*Patient Diagnoses*	*Recommended Drug Therapies*	*Number of Drugs*
				inflammation of eyelid	combination), and other supportive medications	

Decision Tree Multilabel

The Decision Tree algorithm was employed to construct a multilabel classification model that maps input features—patient complaints and diagnoses—to output labels in the form of prescribed drug therapies. The Decision Tree operates by recursively partitioning the data based on the feature that best separates the samples at each node until reaching terminal leaf nodes [32], [33]. In this study, each node in the Decision Tree selects the optimal feature based on the lowest impurity value, resulting in increasingly homogeneous data partitions. The model produces multilabel binary output vectors indicating whether a particular drug therapy is recommended for each patient [34].

$$\hat{Y} = f(X) \tag{1}$$

Gini Impurity

The feature selection criterion used in this study is Gini Impurity, which is the default criterion in the DecisionTreeClassifier implementation of Scikit-learn [35]. Gini Impurity measures the level of data impurity within a node.

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{2}$$

A smaller Gini value indicates greater homogeneity within the node. In the multilabel context, this criterion is applied independently for each label during the splitting process. The resulting decision tree structure maximizes the separation between combinations of complaints and diagnoses with respect to prescribed drug therapies.

Multilabel Evaluation Metrics

To evaluate the performance of the multilabel drug recommendation model, several multilabel evaluation metrics were used to capture different aspects of prediction quality, including overall classification performance, patient-level recommendation accuracy, and label-level prediction errors. The Micro-F1 score was used as the primary evaluation metric because it aggregates true positives (TP), false positives (FP), and false negatives (FN) across all drug therapy labels. This metric is suitable for multilabel datasets with imbalanced label distributions, which are commonly found in medical data. The mathematical formulation of Micro-F1 is presented in Equation (3).

In addition to Micro-F1, the Samples-F1 score was used to evaluate model performance at the patient level. This metric calculates the F1 score for each individual sample by comparing the predicted drug set with the actual therapy set and then averages the results across all samples. Therefore, Samples-F1 directly reflects the model's ability to recommend the correct combination of drugs for each patient. The formulation is shown in Equation (4).

To measure prediction errors, Hamming Loss was used, which calculates the proportion of incorrectly predicted labels relative to the total number of labels. In this study, the metric is expressed as Hamming Score, representing the complement of Hamming Loss. The formulation is provided in Equation (5).

Another metric used is Jaccard Similarity, which measures the overlap between predicted and actual drug therapy labels by comparing the intersection and union of both label sets. This metric indicates how closely the recommended therapies match the actual prescriptions. The formulation is shown in Equation (6).

Finally, Hit@5 was used to evaluate whether at least one relevant drug therapy appears among the top five predicted recommendations. In practical clinical settings, a recommendation system may still be considered useful if it is able to include at least one appropriate therapy among the suggested drugs. The definition of Hit@5 is presented in Equation (7).

Furthermore, recommendation quality was evaluated using Precision@K and Recall@K metrics. Precision@K measures the proportion of relevant drug therapies among the top K recommendations generated by the model. A higher Precision@K value indicates that the model produces more relevant drug recommendations. Conversely, Recall@K measures the proportion of actual drug therapies that are successfully captured within the top K predicted recommendations. In this study, K was set to 5 to align with the Hit@5 evaluation metric. These two metrics complement each other by evaluating both the accuracy and completeness of the recommended drug therapies. Their mathematical formulations are presented in Equations (8) and (9).

Finally, Subset Accuracy was used as a strict evaluation metric for multilabel classification. This metric measures the proportion of samples for which the predicted drug therapy label set exactly matches the true label set. Because this metric requires an exact match of all labels, it is considered a very strict performance measure and often results in relatively low values in multilabel classification tasks. The formulation of Subset Accuracy is shown in Equation (10).

$$Micro\text{-}F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

$$Samples\text{-}F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \mid Y_i \cap \hat{Y}_i \mid}{\mid Y_i \mid + \mid \hat{Y}_i \mid} \tag{4}$$

$$Hamming\ Loss = \frac{1}{N \times L}\sum_{i=1}^{N}\sum_{j=1}^{L}\left|y_{i,j} - \hat{y}_{i,j}\right| \tag{5}$$

$$Jaccard_{samples} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mid Y_i \cap \hat{Y}_i \mid}{\mid Y_i \cup \hat{Y}_i \mid} \tag{6}$$

$$Hit@K = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\left|Y_i \cap \hat{Y}_i\right| > 0\right) \tag{7}$$

$$Precision@K = \frac{1}{N}\sum_{i=1}^{N}\frac{\mid Y_i \cap \hat{Y}_i^{(K)} \mid}{K} \tag{8}$$

$$Recall@K = \frac{1}{N}\sum_{i=1}^{N}\frac{\mid Y_i \cap \hat{Y}_i^{(K)} \mid}{\mid Y_i \mid} \tag{9}$$

$$SubsetAccuracy = \frac{1}{N}\sum_{i=1}^{N} I(Y_i = \hat{Y}_i) \quad (10)$$

Bootstrap Confidence Interval

The bootstrap method is used to estimate the distribution of evaluation metrics by performing random resampling with replacement from the test data.

In this study, the bootstrap process was performed for 1000 iterations, according to the implementation in the research code. In each iteration, a number of samples equal to the size of the test dataset was randomly selected, and multilabel evaluation metrics—such as Micro-F1, Samples-F1, Jaccard Similarity, Hit@5, Hamming Loss, Subset Accuracy, Precision@K, and Recall@K—were calculated based on the model's predictions for that sample. All metric values from each iteration were then collected to form an empirical distribution for each evaluation metric [36].

$$CI_{95\%} = [Q_{0.025}(\theta), Q_{0.975}(\theta)] \quad (11)$$

Where $P_{2.5}$ and $P_{97.5}$ represent the lower and upper bounds of the confidence interval, respectively. The use of the bootstrap confidence interval in this study aims to ensure that the performance of the multilabel Decision Tree model does not depend on a single data split and exhibits good statistical stability.

Result and Discussion

Results

The comparison results presented in **Table 4** show that the Logistic Regression baseline achieves higher scores in several global multilabel evaluation metrics, including Micro-F1, Samples-F1, and Jaccard Similarity. These results indicate that the baseline model is effective in capturing overall label patterns within the dataset. However, the multilabel Decision Tree model still demonstrates reasonably stable performance, with a Micro-F1 value of 0.292 and a Samples-F1 value of 0.281. The Hit@5 value of 0.690 also indicates that in most cases the system is able to include at least one relevant drug within the top recommendations.

Table 4 also highlights that both models produce relatively comparable results across several evaluation metrics. While Logistic Regression achieves slightly higher scores in global multilabel metrics, the Decision Tree model still demonstrates competitive performance. In particular, the model maintains good recommendation capability, as reflected by the Hit@5 value of 0.690 and a relatively low Hamming Loss of 0.079, indicating that prediction errors remain limited. These results suggest that the Decision Tree model is still able to generate relevant drug recommendations despite its simpler structure. Moreover, its interpretable decision rules make it particularly useful for supporting clinical decision-making. Although Logistic Regression achieves slightly higher scores in several global evaluation metrics, the Decision Tree model was selected as the main approach in this study because of its interpretability and transparent decision structure. In clinical decision support systems, the ability to explain how recommendations are generated is an essential requirement to ensure that healthcare professionals can understand and trust the system's outputs.

Table 4. Multilabel Evaluation Results of Decision Tree and Logistic Regression Models

Metric	Decision Tree Multilabel	Logistic Regression
Micro-F1 ↑	0.292	0.412
Samples-F1 ↑	0.281	0.369
Hit@5 ↑	0.690	0.705
Jaccard (Samples) ↑	0.193	0.265
Precision@K ↑	0.218	0.217
Recall@K ↑	0.282	0.286
Subset Accuracy ↑	0.015	0.010
Hamming Loss ↓	0.079	0.045

Despite slightly lower scores in several global metrics, the Decision Tree model shows competitive performance in recommendation-oriented metrics such as Precision@K and Subset Accuracy. More importantly, the Decision Tree provides an interpretable decision structure that allows healthcare practitioners to understand how drug recommendations are generated based on combinations of patient complaints and diagnoses. This interpretability is particularly important in clinical decision support systems, where the reasoning behind a recommendation must be transparent and understandable for medical personnel.

Ablation Study of Feature Configuration

To further examine the contribution of different feature groups, an ablation study was conducted using three input configurations: complaints only, diagnoses only, and the combination of complaints and diagnoses across three dataset scenarios (N=500, N=800, and N=1000). The results are presented in Table 5. From the results, the diagnosis-only configuration generally produces higher performance compared to using complaint features alone. For example, in the N800 dataset scenario, the diagnosis-only configuration achieves a Micro-F1 score of 0.273 and a Samples-F1 score of 0.281, which are the highest among the evaluated configurations. This indicates that diagnosis information provides stronger signals for predicting appropriate drug therapies, since diagnoses represent more specific clinical conditions.

Table 5. Feature Ablation Results

Dataset	*Feature Type*	*Micro -F1*	*Samples-F1*	*Hit@5*	*Jaccard*	*Precision@K*	*Recall @K*	*Subset Accuracy*	*Hamming*
N1000	Diagnoses Only	0.221	0.179	0.450	0.120	0.144	0.174	0.000	0.074
	Complaints + Diagnoses	0.137	0.141	0.384	0.094	0.112	0.140	0.004	0.108
	Complaints Only	0.128	0.117	0.335	0.076	0.088	0.105	0.000	0.090
N500	Diagnoses Only	0.222	0.242	0.550	0.169	0.164	0.213	0.000	0.090
	Complaints + Diagnoses	0.154	0.170	0.432	0.113	0.121	0.165	0.005	0.105
	Complaints Only	0.165	0.167	0.500	0.104	0.126	0.189	0.000	0.100
N800	Diagnoses Only	0.273	0.281	0.688	0.189	0.214	0.277	0.000	0.084
	Complaints + Diagnoses	0.124	0.126	0.314	0.084	0.087	0.110	0.003	0.097
	Complaints Only	0.126	0.127	0.419	0.079	0.098	0.133	0.000	0.100

In contrast, when only complaint features are used, the model tends to produce lower performance across most evaluation metrics. Patient complaints usually describe general symptoms that may appear in different medical conditions, which makes it more difficult for the model to distinguish the appropriate drug therapies. Although the diagnosis-only configuration shows relatively strong performance, combining complaints and diagnoses still provides complementary clinical information. The integration of both features allows the model to capture relationships between patient symptoms and their corresponding diagnoses. This combination helps provide a more complete representation of patient conditions, which is important for generating more meaningful drug recommendations.

Overall, the results of the ablation study indicate that diagnosis features play a major role in predicting drug therapies, while complaint features contribute additional contextual information that can enrich the representation of patient cases in the multilabel classification model.

Performance with Respect to Dataset Size

Figure 2 shows the comparison of model performance across three dataset scenarios (N=500, N=800, and N=1000). Overall, the results indicate that the model maintains relatively stable performance across different dataset sizes, although slight variations can be observed in several metrics. The N500 dataset produces the highest scores for Micro-F1, Samples-F1, Precision@K, and Recall@K, suggesting that the model is able to capture label patterns effectively in this configuration. Meanwhile, the results for N800 and N1000 remain comparable and still demonstrate consistent predictive capability.

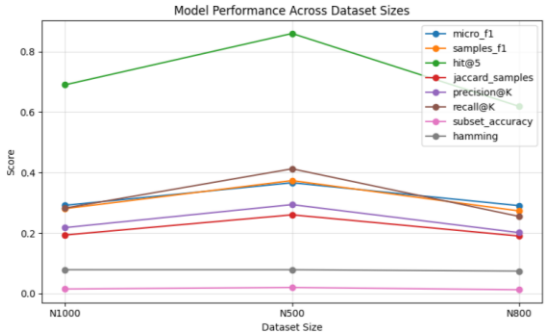

Figure 2. Model Performance Across Different Dataset Sizes

The relatively high Hit@5 values across all scenarios indicate that the model is generally able to include at least one relevant drug therapy within the top five recommendations. Overall, these findings suggest that the multilabel Decision Tree model maintains stable recommendation performance even when the dataset size varies.

Distribution of the Number of Recommended Drugs

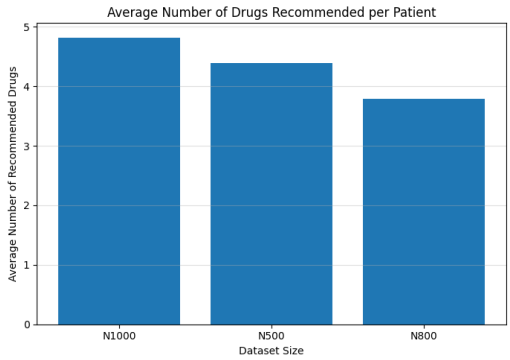

Figure 3. Average Number of Recommended Drugs Across Dataset Sizes

As shown in **Figure 3**, the bar chart illustrates the distribution of the number of recommended drugs across different dataset scenarios. The analysis indicates that the average number of drugs generated by the model is relatively consistent in each dataset scenario. This suggests that the model is not aggressive in recommending drugs and is able to maintain a balance between the number of recommendations and clinical relevance. This consistency, as reflected in the bar chart, is important in the context of real-world implementation, as a drug recommendation system is expected not to generate excessive or overly limited recommendations, but rather to align with common medical practice. In addition, the relatively stable number of recommended drugs across different dataset sizes indicates that the model behavior remains consistent even when the amount of training data changes. This suggests that the model is able to maintain balanced recommendation patterns without producing extreme variations in the number of suggested therapies.

Multilabel Decision Tree

As shown in **Figure 4**, the visualization of the global decision tree illustrates the decision structure formed from patient complaint and diagnosis features. Each node represents the presence condition of a specific complaint or diagnosis, while the leaf nodes represent the probability of recommending a drug therapy. From the structure of the tree, it can be observed that several complaint and diagnosis attributes serve as key splitting criteria that influence the recommended drug therapies. This indicates that the model is able to capture meaningful relationships between patient symptoms, clinical diagnoses, and the corresponding therapeutic treatments.

Furthermore, the hierarchical structure of the tree reflects how the model progressively narrows down the possible drug therapy options based on combinations of complaints and diagnoses observed in the dataset. This hierarchical decision process allows the model to generate multiple therapy recommendations that remain consistent with the patterns found in historical patient records. In addition, the resulting model remains interpretable, allowing medical personnel to understand the basis of the model's decision-making process. This interpretability capability is a primary advantage of the Decision Tree compared to black-box classification methods, particularly in clinical decision support systems.

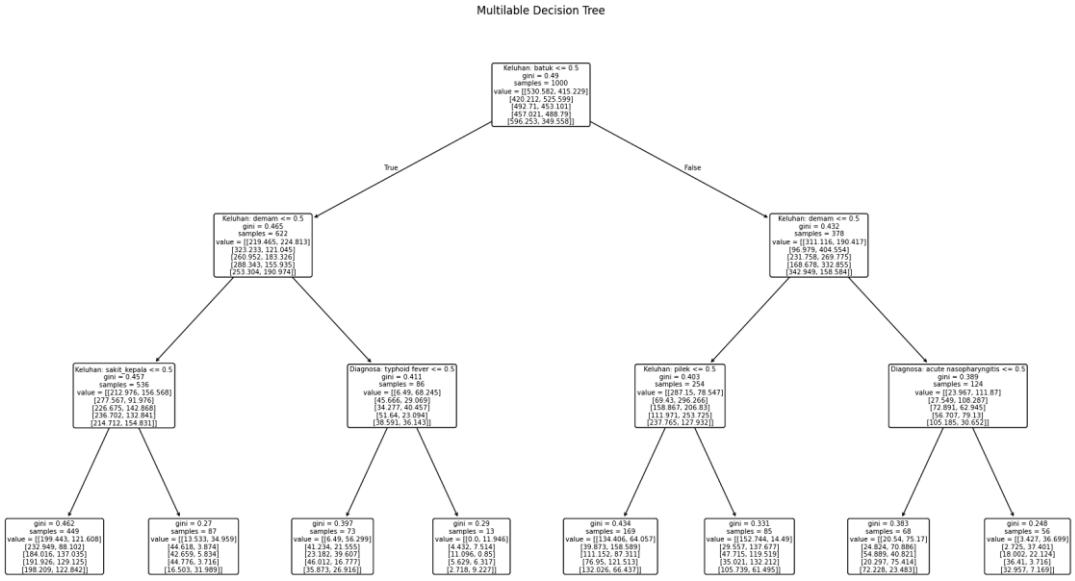

Figure 4. Multilabel Decision Tree

Model Stability Analysis Using Bootstrap

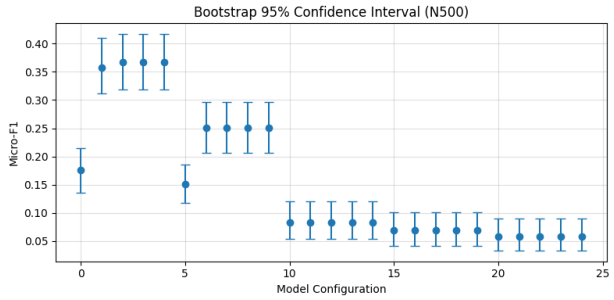

Figure 5. Bootstrap 95% confidence intervals of Micro-F1 across dataset 500

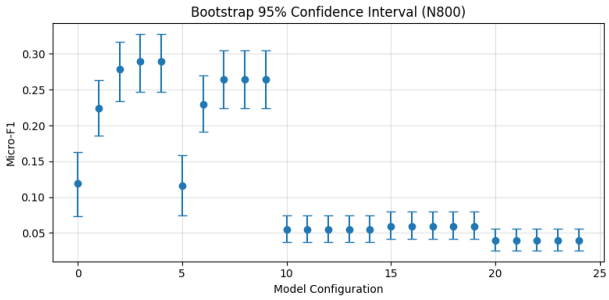

Figure 6. Bootstrap 95% confidence intervals of Micro-F1 across dataset 800

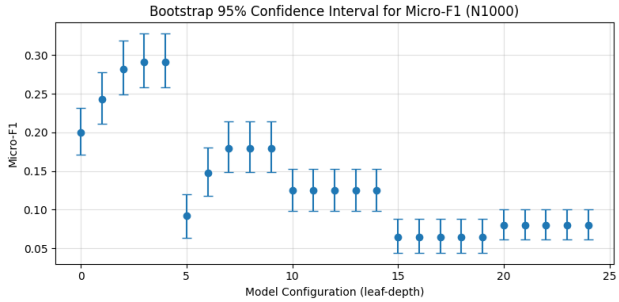

Figure 7. Bootstrap 95% confidence intervals of Micro-F1 across dataset 1000

Figures 5, 6, 7 present the stability analysis results obtained using the bootstrap method with 1000 resampling iterations for three dataset scenarios: N500, N800, and N1000. The analysis evaluates the 95% confidence interval of the Micro-F1 metric across different model configurations.

Overall, the confidence intervals across the three datasets remain relatively narrow, indicating that the model performance is stable under repeated resampling. The N500 scenario shows several configurations with relatively higher Micro-F1 values, while N800 and N1000 demonstrate comparable patterns with consistent interval ranges. The relatively compact distribution of the error bars suggests that the model does not exhibit significant performance fluctuations across different resampled subsets of the data.

Micro-F1 was used in this analysis because it provides a global evaluation by aggregating true positives, false positives, and false negatives across all labels. This makes it suitable for multilabel classification problems with potentially imbalanced label distributions, allowing the bootstrap procedure to assess the stability of the model's overall predictive performance.

Discussion

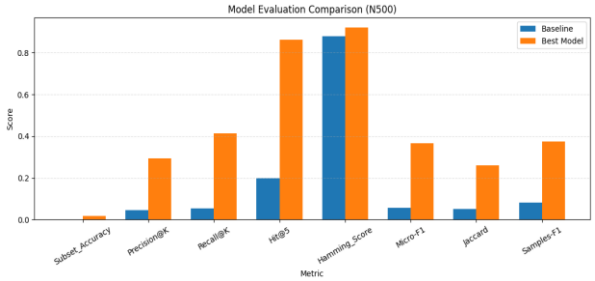

Figure 8. Model Evaluation Comparison Dataset 500

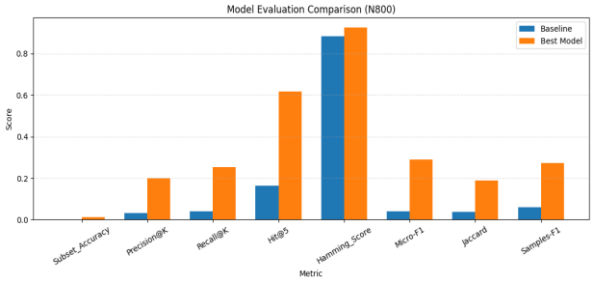

Figure 9. Model Evaluation Comparison Dataset 800

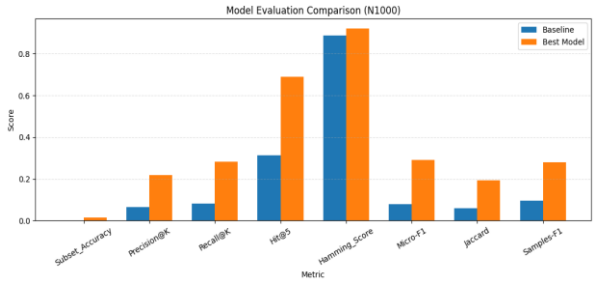

Figure 10. Model Evaluation Comparison Dataset 1000

Figures 8, 9, 10 present a comparison between the baseline configuration and the best-performing model across three dataset scenarios: N500, N800, and N1000. The comparison is evaluated using several multilabel metrics, including Micro-F1, Samples-F1, Jaccard, Precision@K, Recall@K, Hit@5, Subset Accuracy, and Hamming Score. Across all dataset sizes, the best model consistently achieves higher scores than the baseline configuration for most

evaluation metrics. Significant improvements can be observed in Precision@K, Recall@K, Micro-F1, Samples-F1, and Jaccard, indicating that the optimized model is more effective in identifying relevant drug therapy labels and producing more accurate multilabel predictions. In addition, the Hit@5 metric shows a noticeable increase, suggesting that the best model is more likely to include at least one correct drug recommendation within the top five predicted therapies. Meanwhile, the Hamming Score also improves slightly, indicating a reduction in label prediction errors.

Overall, these results demonstrate that the selected hyperparameter configuration improves the model's ability to capture relationships between patient complaints, diagnoses, and drug therapy recommendations. This improvement is consistently observed across the three dataset scenarios, indicating that the optimized multilabel Decision Tree model provides more reliable and accurate recommendation performance.

Error Analysis

An additional error analysis was conducted to examine the types of prediction errors produced by the multilabel Decision Tree model. The analysis focuses on identifying the most frequently missed drug labels (false negatives) and the most frequently over-predicted drug labels (false positives) in the best model configuration under the N=1000 dataset scenario.

Table 6. Error Analysis of the Best Multilabel Decision Tree Model

Dataset	*False Negative Drugs (Missed Predictions)*	*Frequency*	*False Positive Drugs (Over-Predicted)*	*Frequency*
N500	Vitamin C	25	Omeprazole	23
	Paracetamol	22	Aluminium Hydroxide	18
	Prednisone	17	Antacids	18
	Chlorpheniramine	16	Zinc	18
	Acetylcysteine	14	Paracetamol	15
N800	Paracetamol	62	Aluminium Hydroxide	34
	Vitamin C	45	Antacids	34
	Amoxicillin	36	Omeprazole	31
	Chlorpheniramine	29	Magnesium Hydroxide	28
	Vitamin B Complex	23	Cefadroxil	25
N1000	Paracetamol	74	Zinc	31
	Vitamin C	51	Antacids	30
	Vitamin B Complex	37	Aluminium Hydroxide	28
	Chlorpheniramine	37	Domperidone	26
	Amoxicillin	35	Prednisone	23

As shown in **Table 6**, several commonly prescribed drugs such as paracetamol, vitamin C, and vitamin B complex appear among the most frequently missed labels. This behavior may be influenced by the multilabel sparsity of the dataset and the diversity of drug combinations prescribed in clinical practice. On the other hand, drugs such as zinc, antacids, and aluminium hydroxide tend to be over-predicted by the model. These findings indicate that the model occasionally favors general supportive therapies, which appear frequently across different patient cases.

Conclusion

This study developed a drug recommendation system using a multilabel Decision Tree model based on patient complaint and diagnosis data derived from electronic medical records. The multilabel approach allows the model to represent real clinical situations where a patient may receive multiple drug therapies simultaneously while maintaining an interpretable decision structure.

Experimental results show that the model achieves relatively stable performance across different dataset scenarios (N=500, N=800, and N=1000). Although the Logistic Regression baseline achieves slightly higher scores in several

global multilabel metrics, the Decision Tree model remains competitive while offering an important advantage in interpretability for clinical decision support. The ablation study indicates that diagnosis features provide stronger predictive signals for drug therapy recommendations, while complaint features contribute complementary contextual information. In addition, bootstrap validation confirms that the model performance is statistically stable, and the optimized model configuration consistently improves recommendation performance across multiple evaluation metrics.

Overall, this study contributes to the development of an interpretable multilabel drug recommendation framework that integrates patient complaints and diagnoses to support data-driven clinical decision-making.

However, this study has several limitations. The dataset was obtained from a single healthcare facility, which may limit the generalizability of the model to other clinical environments. Future research may involve incorporating data from multiple healthcare institutions and exploring additional machine learning approaches to further improve the robustness and applicability of drug recommendation systems in real-world healthcare settings.

Acknowledgment

The authors would like to express their sincere gratitude to the Community Health Center in Pangkajene and Kepulauan Regency (Pangkep) for the permission, support, and cooperation provided during the process of collecting and providing research data. This support greatly facilitated the implementation of the study, particularly in obtaining relevant and structured medical record data, enabling this research to be completed successfully.

References:

[1] J. Wang, X. Chen, and Y. Li, "Structure Design and Optimization Algorithm of a Lightweight Drive Rod for Precision Die-Cutting Machine," *Applied Sciences (Switzerland)*, vol. 13, no. 7, Apr. 2023, doi: 10.3390/app13074211.

[2] A. Raţiu and E.-L. Pop, "Machine Learning in Clinical Decision Making: Applications, Data Limitations and Multidisciplinary Perspectives," *Applied Sciences*, vol. 16, no. 2, p. 785, Jan. 2026, doi: 10.3390/app16020785.

[3] X. Yao, A. Rao, and R. Padman, "Analytical approaches for medication management for patient safety: a scoping review," *npj Health Systems*, vol. 2, no. 1, Dec. 2025, doi: 10.1038/s44401-025-00052-1.

[4] S. E. M. Purba, "A Comparative Study of Drug Prediction Models using KNN, SVM, and Random Forest," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 378–392, Mar. 2025, doi: 10.51519/journalisi.v7i1.1013.

[5] Y. Tang *et al.*, "LAMRec: Label-aware Multi-view Drug Recommendation," in *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, Oct. 2024, pp. 2230–2239. doi: 10.1145/3627673.3679656.

[6] H. Darwis, F. A. Syahrir, and L. N. Hayati, "A Hybrid Movie Recommendation System to Address Data Sparsity Using Genre-Based K-Means and Neural Collaborative Filtering," *ILKOM Jurnal Ilmiah*, vol. 17, no. 2, pp. 203–212, Sep. 2025, doi: 10.33096/ilkom.v17i2.2868.203-212.

[7] A. Putri, D. Azka Faz, and F. T. Hafizhulloh, "Classification of Drug Types using Decision Tree Algorithm," *Journal of Dinda Data Science, Information Technology, and Data Analytics*, vol. 3, no. 2, pp. 65–70, 2023.

[8] L. Zhou, X. Zheng, D. Yang, Y. Wang, X. Bai, and X. Ye, "Application of multi-label classification models for the diagnosis of diabetic complications," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01525-7.

[9] F. Kamran *et al.*, "Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: Model development and multisite external validation study," *The BMJ*, vol. 376, Feb. 2022, doi: 10.1136/bmj-2021-068576.

[10] E. Johns *et al.*, "Using machine learning to predict pharmaceutical interventions during medication prescription review in a hospital setting," *American Journal of Health-System Pharmacy*, vol. 82, no. 22, pp. 1238–1248, Nov. 2025, doi: 10.1093/ajhp/zxaf089.

[11] Y. Salim, A. P. Utami, A. R. Manga, H. Azis, and F. T. Admojo, "Optimal Strategy for Handling Unbalanced Medical Datasets: Performance Evaluation of K-NN Algorithm Using Sampling Techniques," *Knowledge Engineering and Data Science*, vol. 7, no. 2, Dec. 2024, doi: 10.17977/um018v7i22024p176-186.

[12] P. Lestari, L. Belluano, R. A. Rahma, H. Darwis, and A. R. Manga, "Analysis of ensemble machine learning classification comparison on the skin cancer MNIST dataset," *Computer Science and Information Technologies*, vol. 5, no. 3, pp. 235–242, 2024, doi: 10.11591/csit.v5i3.pp235-242.

[13] Purnawansyah, A. P. Wibawa, T. Widiyaningtyas, Haviluddin, H. Darwis, and H. Azis, "An in-depth exploration of supervised and semi-supervised learning on face recognition," *Open Computer Science*, vol. 15, no. 1, Jan. 2025, doi: 10.1515/comp-2025-0029.

[14] J. L. Montalvo, J. C. Silva, and A. Zamora-Mendez, "TKEO-DESA-Based decision tree for power quality events detection and classification," *Electric Power Systems Research*, vol. 252, Jan. 2026, doi: 10.1016/j.epsr.2025.112387.

[15] Dewi Widyawati and Amaliah Faradibah, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indonesian Journal of Data and Science*, vol. 4, no. 2, pp. 80–89, Jul. 2023, doi: 10.56705/ijodas.v4i2.76.

[16] O. Khalaf, A. Ben Ishak, and S. García, "Towards explainable multilabel learning: Fusing label dependency analysis with monotonic decision trees," *Information Fusion*, vol. 127, Mar. 2026, doi: 10.1016/j.inffus.2025.103691.

[17] F. Liu, W. Wang, J. Zheng, Y. Xie, X. Wang, and D. Zhang, "EDRMM: enhancing drug recommendation via multi-granularity and multi-attribute representation," *BMC Bioinformatics*, vol. 26, no. 1, Dec. 2025, doi: 10.1186/s12859-025-06167-4.

[18] G. Liu *et al.*, "DNMDR: Dynamic networks and multi-view drug representations for safe medication recommendation," *Knowl. Based. Syst.*, vol. 329, Nov. 2025, doi: 10.1016/j.knosys.2025.114327.

[19] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *Expert Syst. Appl.*, vol. 203, Oct. 2022, doi: 10.1016/j.eswa.2022.117215.

[20] W. T. Kim *et al.*, "Medication Extraction and Drug Interaction Chatbot: Generative Pretrained Transformer-Powered Chatbot for Drug-Drug Interaction," *Mayo Clinic Proceedings: Digital Health*, vol. 2, no. 4, pp. 611–619, Dec. 2024, doi: 10.1016/j.mcpdig.2024.09.001.

[21] J. Khatib Sulaiman Dalam No, H. Akram Abdulqader, and A. Mohsin Abdulazeez, "A Review on Decision Tree Algorithm in Healthcare Applications," *Indonesian Journal of Computer Science*.

[22] K. Chen, M. Ao, S. Moon, G. Burns, and Q. Zhu, "Machine learning-based identification of natural history studies in rare diseases: a step toward understanding disease development and outcome," *Journal of Rare Diseases (Germany)*, vol. 4, no. 1, Dec. 2025, doi: 10.1007/s44162-025-00115-9.

[23] S.-K. Tan, S.-C. Chong, K.-K. Wee, and L.-Y. Chong, "Personalized Healthcare: A Comprehensive Approach for Symptom Diagnosis and Hospital Recommendations Using AI and Location Services," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 117–135, Feb. 2024, doi: 10.33093/jiwe.2024.3.1.8.

[24] R. Su, H. Yang, L. Wei, S. Chen, and Q. Zou, "A multi-label learning model for predicting drug-induced pathology in multi-organ based on toxicogenomics data," *PLoS Comput. Biol.*, vol. 18, no. 9 September, Sep. 2022, doi: 10.1371/journal.pcbi.1010402.

[25] L. Y. Jiang *et al.*, "Health system-scale language models are all-purpose prediction engines," *Nature*, vol. 619, no. 7969, pp. 357–362, Jul. 2023, doi: 10.1038/s41586-023-06160-y.

[26] F. Sogandi, "Identifying diseases symptoms and general rules using supervised and unsupervised machine learning," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-69029-8.

[27] T.-T. Nguyen *et al.*, "Mimic-IV-ICD: A new benchmark for eXtreme MultiLabel Classification," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.13998

[28] A. Garchitorena *et al.*, "Expanding community case management of malaria to all ages can improve universal access to malaria diagnosis and treatment: results from a cluster randomized trial in Madagascar," *BMC Med.*, vol. 22, no. 1, Dec. 2024, doi: 10.1186/s12916-024-03441-9.

[29] Z. Wang *et al.*, "ICDXML: enhancing ICD coding with probabilistic label trees and dynamic semantic representations," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-69214-9.

[30] M. Arjmandi, M. Fattahi, M. Motevassel, and H. Rezaveisi, "Evaluating algorithms of decision tree, support vector machine and regression for anode side catalyst data in proton exchange membrane water electrolysis," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-47174-w.

[31] Z. Ali *et al.*, "Deep Learning for Medication Recommendation: A Systematic Survey," Mar. 01, 2023, *MIT Press Journals*. doi: 10.1162/dint_a_00197.

[32] K. ei Sada *et al.*, "Development and validation of data-driven, decision tree–based algorithms for identifying Behçet's disease in claims data," *Int. J. Med. Inform.*, vol. 209, Apr. 2026, doi: 10.1016/j.ijmedinf.2026.106266.

[33] S. Rahmah Jabir, H. Azis, and S. H. Mansyur, "Enhancing The Quality of College Decisions Through Decision Tree and Random Forest Models."

[34] X. Zhu *et al.*, "Escitalopram treatment for patients with major depressive disorder: decision trees for treatment algorithm," *J. Psychiatr. Res.*, vol. 195, pp. 284–290, Apr. 2026, doi: 10.1016/j.jpsychires.2026.02.001.

[35] M. A. Bouke, A. Abdullah, S. H. ALshatebi, and M. T. Abdullah, "E2IDS: An Enhanced Intelligent Intrusion Detection System Based On Decision Tree Algorithm," *Journal of Applied Artificial Intelligence*, vol. 3, no. 1, pp. 1–16, Jun. 2022, doi: 10.48185/jaai.v3i1.450.

[36] F. Wang, J. Chu, L. Shen, and S. Chang, "MESM: integrating multi-source data for high-accuracy protein-protein interactions prediction through multimodal language models," *BMC Biol.*, vol. 23, no. 1, Dec. 2025, doi: 10.1186/s12915-025-02356-y.