Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 2087-2098
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Automated Drug Name Mapping Using Weighted Ensemble Natural Language Processing and Machine Learning Reranker

Azmi Fadhilah Hasibuan[1*], Shumaya Resty Ramadhani[2], Akhas Rahmadeyan[3]

[1,2]Jurusan Teknologi Informasi, Politeknik Caltex Riau, Indonesia
[3]Department of Data Analytics, GUE Ecosystem, Indonesia

E-Mail: [1]azmi22si@mahasiswa.pcr.ac.id, [2]shumaya@pcr.ac.id, [3]akhas.rahmadeyan@assist.id

Received Apr 14th 2026; Revised Jun 22th 2026; Accepted Aug 30th 2026; Available Online Oct 29th 2026
Corresponding Author: Azmi Fadhilah Hasibuan
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Drug names in Electronic Medical Records (EMR) are often entered by healthcare workers as free text, resulting in spelling errors, inconsistent formatting, and ambiguous abbreviations that hinder automated mapping to standardized drug databases. This study proposes an automated drug name-mapping system that integrates multi-method ensemble scoring with a supervised machine learning reranker. The system employs Term Frequency–Inverse Document Frequency (TF-IDF) character n-grams for candidate retrieval, followed by a weighted ensemble of five similarity methods, including Best Matching 25 (BM25) and fuzzy string-matching variants, supplemented by pharmaceutical-domain features. A Random Forest reranker is subsequently applied to improve ranking accuracy. Experiments were conducted on 9,326 free-text prescriptions against a master database of 55,078 drug entries, using an 80:20 stratified split to train and evaluate the reranker. The proposed system achieved a Top-1 Accuracy of 79.53%, representing an 18.53 percentage-point improvement over the Euclidean Distance baseline (61.00%). The confidence-scoring mechanism categorized predictions into three tiers: HIGH-confidence cases achieved 87.44% accuracy, and LOW-confidence cases achieved 28.30% accuracy, enabling selective automation for HIGH-confidence cases and targeted manual review for ambiguous cases. These results confirm that ensemble-based Natural Language Processing (NLP), combined with machine-learning-based reranking, provides a reliable solution for automated drug name standardization in clinical information systems.

Keywords: *Drug Name Mapping, Ensemble Scoring, Fuzzy String Matching, Machine Learning Reranker, NLP.*

1. INTRODUCTION

In today's era of globalization, the need for integrated information technology in the healthcare sector is crucial. Technological advances, particularly the internet, have made it increasingly easier for people to access healthcare services. In Indonesia, the Minister of Health Regulation Number 24 of 2022 requires all healthcare facilities to maintain Electronic Medical Records (EMR) [1]. This policy requires data integration between healthcare facilities and the SATUSEHAT platform.

In this system, prescription data are still entered manually by medical personnel in free text. This manual prescription data entry process can result in inconsistent drug data, irregular formatting, and typos. Medication errors are inherently high-risk due to the numerous influencing factors, such as brand name, ingredients, dosage, and type. This irregularity in drug names prevents the system from mapping prescriptions to standard medical terminology, thereby increasing the risk of medication errors [2]. Globally, medication-related errors are estimated to occur in 1 in 20 patients, and more than half of these incidents (53%) begin at the prescribing stage [3]. According to a 2023 study, up to 91% of medication errors are due to prescribing errors [4]. In Indonesia, a retrospective study at a large hospital found that prescribing-stage medication errors were significant, with key factors including manual prescribing and the lack of a standardized e-prescribing system [5]. The irregularity in entry formats is not merely an administrative issue, but a real threat to patient safety in healthcare facilities. Recent research shows that implementing electronic medical records without strict data-entry standardization can lead to medication errors and duplicate therapies, compromising clinical safety [6][7]. Therefore, a systematic solution is needed in the form of an automatic text-matching algorithm to bridge the gap between medical personnel input and standard master data.

This study employed Natural Language Processing (NLP) methods that combined several token-based string-matching algorithms. Term Frequency–Inverse Document Frequency (TF-IDF) n-grams is a text feature weighting technique that has proven effective for improving text-matching performance across various

domains, including classification and information retrieval [8]. In the context of medical and pharmaceutical data, Best Matching 25 (BM25) is an effective component in hybrid retrieval systems, including in a study of Indonesian drug data retrieval [9]. This method was chosen because drug names in unstructured prescription data are often written in an inconsistent order, including the generic name, manufacturer name, dosage, and drug form.

While recent advancements in Transformer-based architectures have shown promise in NLP, this study intentionally adopts a Random Forest-based reranking approach. This decision is driven by the need for high interpretability in clinical decision-making and the computational efficiency required for deployment in resource-constrained healthcare facilities, where real-time performance is critical without high-end GPU clusters. Random Forest, a decision-tree-based ensemble method, has been widely used for medical data classification and has demonstrated competitive performance with interpretability through feature importance analysis [10].

Although several studies have explored drug name normalization using ontology-based approaches such as RxNorm [11] or transformer-based medical concept modeling [12], their application to Indonesian prescription data is still very limited. Previous studies in the context of Indonesian pharmaceutical data generally focus on language model-based retrieval, such as MPNet, without integrating supervised learning-based reranking mechanisms [9]. This study fills this gap by proposing a drug name mapping system that combines a multi-method scoring ensemble with a supervised-trained Random Forest-based reranker specifically designed for the characteristics of free-form prescription text in the context of Indonesian EMRs.

The process of matching over-the-counter prescription texts to standard master data is often hampered by numerous spelling variations and typos [13]. To address this issue, modern information retrieval architectures have begun implementing a two-stage retrieval-and-rerank system. The first stage focuses on search speed to narrow down candidate data (the retriever), while the second stage refines the results to find the most precise matches (the reranker) [14]. Text-based search methods such as Term Frequency–Inverse Document Frequency (TF-IDF) and Best Matching 25 (BM25) are highly efficient at narrowing the initial search space but often fail to handle inconsistently written local abbreviations [15]. In Indonesia, the trend in NLP research on medical records is shifting toward the use of Large Language Models (LLMs), but this approach requires significant and expensive computing resources [16]. This research fills this research gap. No study has specifically combined the flexibility of multiple text-matching algorithms (a weighted ensemble of TF-IDF, Best Matching 25 (BM25), and fuzzy strings) with the Random Forest algorithm as a machine-learning reranker for local drug prescription standardization. This hybrid approach offers an accurate yet lightweight search solution, enabling smooth execution without a high-spec GPU cluster.

By using this combination of methods, the most effective prescription text-mapping model can be identified. To further improve accuracy, this study applies to a Machine Learning Reranker with Random Forest and Extreme Gradient Boosting (XGBoost) algorithms, trained on features derived from ensemble scoring. This approach is not only useful for improving data quality and the efficiency of internal data analysis teams, but also for ensuring consistency in drug prescription management. Ultimately, this modern analytical approach has great potential to improve overall healthcare services by ensuring that the data used is structured, reliable, and meets national health data connectivity standards on the SATUSEHAT platform.

2. MATERIALS AND METHOD

This research method consists of seven stages, as shown in the figure, designed to produce an optimal model for mapping free prescription text to drug master data. The overall workflow is presented in Figure 1.

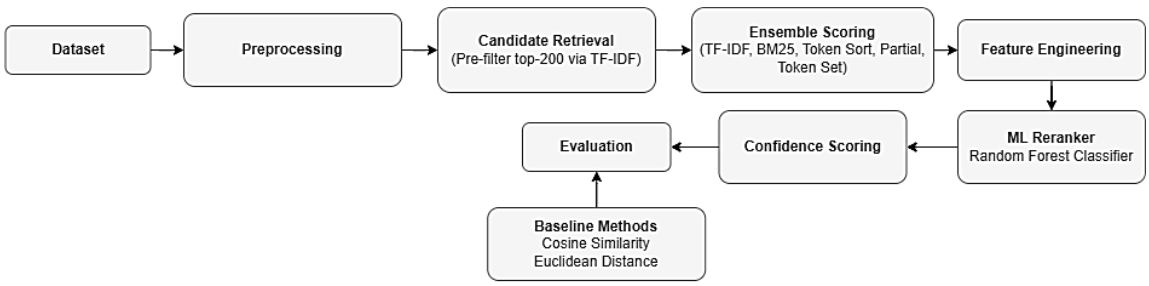


Figure 1. Workflow for Drug Name Mapping Using an Ensemble Approach and ML Reranker

2.1. Literature Review

2.1.1. Text Mining

Text Mining is the process of converting unstructured text into a more structured format to identify important patterns. This process is used to analyze large collections of text to uncover core concepts, trends, and hidden patterns within them [17]. Text Mining is also useful for quickly analyzing information, such as

sentence sentiment, making it easier for users to obtain relevant, high-quality data [18]. Text Mining generally includes several main stages, such as tokenization (dividing text into words), stopword removal, stemming, and calculating word weights using methods such as Term Frequency–Inverse Document Frequency (TF-IDF).

2.1.2. TF-IDF & N-gram

The Term Frequency-Inverse Document Frequency (TF-IDF) method converts text into numerical data, enabling each word to be weighted. TF-IDF is a statistical tool used to assess the significance of a word in a document [19]. This technique combines two concepts to calculate the value, namely how often a word appears in a particular document and how many documents do not contain that word. The frequency of a word's appearance in the document in question describes the level of importance of that word in that document [20]. In the context of information retrieval, TF-IDF is very important for determining keywords that differentiate between documents: a high TF value in a document, combined with a low IDF, indicates that the word is general and provides little information. Conversely, a high IDF value indicates that the word is more specific and provides more useful information [21].

2.1.3. BM25

BM25 (Best Match 25) is a probabilistic ranking function that extends the TF-IDF principle by adding term frequency saturation and document length normalization. Unlike conventional TF-IDF, BM25 limits the contribution of excessive term frequencies, making it more robust to variations in text length [22]. In the biomedical and pharmaceutical domains, BM25 remains a competitive baseline due to its ability to handle precise lexical matching, especially for named entities such as drug names [22].

2.1.4. Fuzzy String Matching

Fuzzy string matching is a technique that flexibly measures the similarity between two strings, without requiring an exact match. This method includes variants such as Token Sort Ratio, Partial Ratio, and Token Set Ratio, each designed to handle word-order inconsistencies, string-length differences, and token redundancy [23]. This approach has proven effective in domains where free text contains high spelling variation [23].

2.1.5. Euclidean Distance

Euclidean Distance is one of the most common methods for measuring distance and determining the similarity between two vectors in a multidimensional space. This method works by calculating the square root of the sum of squared differences in the coordinates of the corresponding points. In text processing, the resulting distance value is inversely related to the level of similarity. The lower the distance value, the greater the similarity between two document vectors [24].

2.1.6. Cosine Similarity

Cosine similarity is a fundamental measure widely used in data mining. In general, the Cosine Similarity method measures the angle between two objects (e.g., documents 1 and 2) represented as vectors, using keywords from each document as a benchmark. The core of this method is to assess directional similarity rather than absolute distance, so that the resulting value falls within the range 0 to 1 [25].

2.1.7. XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient-boosting-based ensemble algorithm that builds a predictive model incrementally, with each new decision tree trained to correct the errors of the previous one. XGBoost uses a sequential approach, optimizing an objective function with a second-order gradient technique, resulting in faster convergence and higher accuracy on tabular data. In the context of EHR-based clinical data, XGBoost and Random Forest are the two most widely used machine learning algorithms, with XGBoost being the third most used algorithm in electronic health data-based prediction studies [26].

2.1.8. Random Forest

Random Forest is a machine learning method that consists of a collection of decision trees arranged to create a more precise model. Random Forest creates multiple trees using sample data, with the trees formed during the training process unaffected by previous trees. Furthermore, decisions are made based on majority vote [27].

2.1.9. Ensemble Scoring

Ensemble scoring combines scores from multiple measurement methods to produce more stable and accurate predictions than any single method [28]. The basic principle is that the weaknesses of one method can be compensated for by the strengths of another, resulting in a combined result with lower variance and less bias. In the domains of information retrieval and text matching, ensemble approaches have consistently improved performance, including in clinical entity matching tasks with high spelling variation [23].

2.2. Proposed Drug Mapping Strategy

The initial process involved transforming text using TF-IDF character n-grams to rapidly pre-filter candidates from over 50.000 primary data sets to the top 200 for each query. To address inconsistent word order, this study employed three variants of token-based fuzzy string matching:

1. Token Sort Ratio: Resolves word-order discrepancies by sorting tokens alphabetically before comparisons.
2. Partial Ratio: Designed to handle situations where the drug name in the prescription is shorter than the name in the main data.
3. Token Set Ratio: Addresses spelling variations with repeated or redundant words.

These five methods (TF-IDF, BM25, and the three fuzzy variants) are combined in a weighted ensemble scheme, along with additional features from the pharmaceutical domain, such as dose, manufacturer, and packaging match bonuses.

2.3. Dataset Collection and Compilation

The dataset was obtained from a commercial electronic medical records information system used by a healthcare facility in Indonesia through a research collaboration with the platform developer. The input data consisted of 9.326 drug names from free-text prescriptions, supplemented with ground-truth data in the form of prescription names corresponding to the master prescription data. The first stage was dataset collection, consisting of:

1. Free-text prescription data inputted by medical personnel.
2. Standardized official drug master data.
3. Ground truth data in the form of validation results for matching prescription names.

The dataset description is presented in Table 1.

Table 1. Dataset Description

Dataset Description	Total Data
Total free-text drug prescription data	9.326
Total official drug master data	55.078
Total ground truth data	9.326

2.4. Data Preprocessing

The preprocessing stage in this study aims to standardize all drug name texts before the matching process. This process is applied consistently to both input and master data, ensuring that all datasets have a uniform representation before comparison. This process is important because formatting differences can affect similarity scores and reduce system accuracy. There are three steps carried out in this preprocessing:

1. Basic normalization: All text is converted to lowercase, removing empty () characters or those containing administrative codes that do not contribute to drug identity. Punctuation, periods, excess spaces, and other special characters are removed.
2. Pharmaceutical abbreviation normalization dictionary: created based on an analysis of the most frequently encountered spelling variation patterns in the drug prescription dataset. The pattern identification process was performed manually by sampling the input data, which generated several normalization rules in three broad categories: drug dosage form ("tab" becomes "tablet"), packaging notation (@50 becomes "box 50"), and variations in drug manufacturer names.
3. Generic drug name normalization dictionary: created based on an analysis of discrepancy patterns between master data and prescription data. It comprised two main categories: commonly used generic-name abbreviations (e.g., PCT for paracetamol) and brand names mapped to official generic names.
4. Final check: After all normalizations were applied, remaining non-alphanumeric characters were removed, and whitespace normalization was performed to ensure the text data was clean and ready for the next stage.

2.5. Text Representation and Candidate Retrieval

Given the large size of the master data, which reaches 55.078 entries, directly matching all master data for each query would be highly inefficient. Therefore, a candidate retrieval stage was implemented to refine the initial options before proceeding to a more detailed scoring process. In this stage, all master data was converted into numeric vectors using the TF-IDF method with a character n-gram approach. The character n-gram approach was chosen because it captures character-level similarities, making it more resilient to spelling variations, typos, and abbreviations commonly found in free-text drug prescription data. Each query from the prescription data was then converted into the same vector representation, and the cosine similarity between the query vector and all master vectors was calculated to identify the top 200 candidates. This stage reduces the

search space from 55,078 entries to 200 candidates per query, achieving a Recall@200 of 98.36% across the entire dataset and a processing time approximately 15-20 times faster compared to exhaustive matching, confirming that the 200-candidate threshold maintains sufficient coverage for subsequent assessment stages.

2.6. Ensemble Scoring

Ensemble scoring was implemented in this study by combining scores from five text similarity methods with weights. These five main components run in parallel and independently. The first utilizes TF-IDF Cosine Similarity with a character n-gram approach. The second component uses BM25, which computes relevance scores directly from word frequencies and document length normalization. In this case, BM25 eliminates the need to convert text into vectors, as with TF-IDF. The next three components are fuzzy string matching, such as token sort ratio, partial ratio, and token set ratio, which directly compare string similarity at the token (single-word) level.

This method aims to produce a final score that is more accurate and stable than using a single method, as the advantages of other methods can offset their weaknesses. From the 200 selected candidates, similarity measurements were performed using the weighted ensemble method. Each method was assigned a weight proportional to its expected discriminative contribution to the drug prescription matching task, as detailed in the Rationale column of Table 2:

Table 2. Ensemble Scoring Methods and Weights Used

Method	Weight	Rationale
TF-IDF Cosine Similarity	0.35	Character-based vector representation with high discriminative power.
BM25	0.20	Probabilistic ranking based on token relevance.
Token Sort Ratio	0.25	Robust against word order differences.
Partial Ratio	0.10	Robust against drug names shorter than master data entries.
Token Set Rasio	0.10	Robust against repeated or redundant words.

In addition to the five main scores, additional rule-based pharmaceutical features were added to improve the ability to distinguish similar drug names despite having different attributes:

1. Dosage bonus (+0.10): awarded if the dosage information in the query (e.g., "10mg") matches the candidate.
2. Manufacturer bonus (+0.10): awarded if the manufacturer name in the query matches the candidate.
3. Packaging bonus (+0.15): awarded if the packaging information in the "@number" format in the query matches the candidate.

The packaging bonus is set higher (+0.15) than the dose and manufacturer bonuses (+0.10) because packaging is the most specific attribute in differentiating master entries: two drugs with identical generic names, doses, and manufacturers but different packaging (e.g., box 50 and box 100) are products that need to be clinically differentiated. These three bonuses are additive to the base ensemble score and are only active if the corresponding information is present in the input text. Hence, their contribution to the final score depends on the completeness of the data entered by the healthcare professional. The result of this stage is a prediction table containing the best candidate (Top-1), the second-place candidate (Top-2), and the total score for each query.

2.7. Feature Engineering

After the ensemble scoring process produces results, feature extraction is performed to prepare the input for the machine learning model in the next stage. A total of 10 features are extracted and divided into three categories. The feature descriptions are presented in Table 3.

Table 3. Ensemble Score Features

Ensemble Score Features	
skor_ensemble	Weighted combined score from all methods.
skor_tfidf	Individual TF-IDF cosine similarity score.
skor_bm25	Individual BM25 score.
skor_fuzzy	Combined score from Token Sort, Partial, and Token Set Ratio.
gap_top1_top2	The difference between the ensemble scores of the first candidate and the second candidate

The additional feature gap_top1_top2 is defined as the difference between the ensemble scores of the top-ranked and second-ranked candidates. This margin reflects the system's relative certainty in selecting the best match compared to the closest alternatives.

Table 4. Domain Bonus Features

Domain Bonus Features	
bonus_dosis	Value of 0.10 if dosage matches, 0 otherwise.
bonus_produsen	Value of 0.10 if the manufacturer matches, 0 otherwise.

Table 5. Structural features

Structural Features	
panjang_query	Character length of the drug name in the query.
panjang_kandidat	Character length of the drug name in the master candidate.
selisih_panjang	Difference between panjang_query and panjang_kandidat.

2.8. Machine Learning Reranker

In this stage, the machine learning model is trained on 10 features generated via feature engineering to capture non-linear patterns in feature combinations that a simple weighted ensemble approach cannot capture. Two algorithms were used and compared: Random Forest and XGBoost. To ensure method replication, the total dataset was randomly divided using a stratified random sampling technique into 80% training data (7,460 samples) for training the reranker architecture and 20% testing data (1,866 samples) for final performance testing. This implicit randomization was performed to maintain a balanced distribution of bonus features in the pharmaceutical domain across both data subsets. Random Forest was trained with 200 decision trees (n_estimators = 200), where each tree cast a vote to determine the correct prediction, and the majority vote across all trees determined the final decision. In addition to prediction, the Random Forest model also provides a feature of importance analysis, indicating how much each feature contributes to model accuracy.

As an additional comparison, baseline Euclidean Distance and Cosine Similarity methods, without going through the ensemble and reranker stages, were also evaluated to gauge the extent of improvement provided by the proposed system.

2.9. Evaluation

The selection of evaluation metrics in this study was based on the task at hand, namely, information retrieval. In this case, the system must find one correct drug name from approximately 50.000 candidates for each query. System performance is evaluated using the four metrics in Table 6 as follows:

Table 6. Evaluation

Evaluation Type	Description
Top-1 Accuracy	The percentage of cases in which the system's first prediction exactly matches the ground truth. This metric indicates how accurate the system is in fully automated mode [29].
Top-2 Accuracy	The percentage of cases in which the correct answer appears among the top two predictions. This metric is important for semi-automatic usage scenarios that require confirmation from the user [29].
Mean Reciprocal Rank (MRR)	The average of the reciprocal ranks of the correct answer's position among all candidates. An MRR value approaching 1.0 indicates that the system consistently places the correct answer at the top position. MRR can distinguish two situations that Top-1 Accuracy considers equivalent [30].
Confidence-Level Evaluation	This evaluation is necessary because the system not only produces predictions but also generates a confidence score for each prediction. The confidence score is used to verify that higher confidence levels correspond to predictions that can be accepted without further review. System accuracy is evaluated across three confidence categories: HIGH (score $\geq$ 0.70), MEDIUM (0.50 $\leq$ score < 0.70), and LOW (score < 0.50) [12].

3. RESULTS AND DISCUSSION

3.1. Preprocessing Results

In this study, the preprocessing stage successfully standardized all input and master data into a uniform text representation before the matching process. The preprocessing results for several data samples are presented in Table 7.

Table 7 shows that each sample underwent at least one type of transformation, ranging from expansion of abbreviations and packaging notation conversion to normalization of manufacturer and generic names. This process is necessary because, without it, the texts "GG 100MG TAB @100 TRIFA" and "GUAIFENESIN 100MG TABLET BOX 100 TRIFA" would be treated as two distinct text entities by the system, even though they both refer to the same prescription drug.

Table 7. Preprocessing Steps

Input Text (Before Preprocessing)	Output Text (After Preprocessing)	Transformation
(BPJS) AMLODIPINE 10 MG TABLET @50 INTERBAT. TABLET	amlodipine 10 mg tablet box 50 interbat tablet	Remove (BPJS), @50→box 50, remove period
GG 100 MG TAB @100 TRIFA	guaifenesin 100mg tablet box 100 trifa	GG→guaifenesin, tab→tablet, @100→box 100
PCT 500MG TABLET	Paracetamol 500mg tablet	PCT→paracetamol

The pharmaceutical abbreviation normalization dictionary proved effective in addressing three groups of spelling variations most frequently found in prescription data. The first group is dosage form variations, where abbreviations such as "tab." "tabs," and "TAB" were successfully normalized to "tablet." The second group is packaging notation, where the conversion of the "@number" format to "box number" is one of the most important transformations because this format is used only in input data but never appears in master data, so without this normalization, the system would never be able to match packaging information correctly. The third group is manufacturer name variations, where the spellings "HEXPHARM JAYA" and "HEXPHARM" referring to the same manufacturer were successfully standardized.

The generic drug name normalization dictionary addresses a different but equally important issue: the gap between the abbreviations or brand-name spellings in prescription data and the official generic-name spellings in master data. Without this normalization, abbreviations such as "GG" and "PCT" would never match the full name in master data, even though they refer to the same drug.

Analysis of the entire dataset revealed that only 2.3% of the total data (211 of 9.326 rows) contained explicit packaging information in the "@number" format. This indicates that most healthcare professionals do not consistently include packaging information when inputting prescriptions. Consequently, the packaging bonus feature added during the ensemble scoring stage was active only on these 211 records, thereby limiting its contribution to overall accuracy improvement. However, for the subset of data that did contain packaging information, this normalization played a key role in distinguishing master entries with identical generic names and manufacturers but different packaging quantities.

3.2. Candidate Retrieval Results

The candidate retrieval stage in this study helps improve the efficiency of the matching process, given the large size of the master data. The matching process was performed on approximately 55.078 master data sets for each query from 9.326 input data sets, meaning the system had to perform similarity calculations on 55.078 × 9.326, or over 513 million text pairs. This imposes a computational burden that is impractical to implement directly, especially when the complex fuzzy string-matching method is applied to the entire data set.

TF-IDF character n-grams were used as an initial filter to represent the text into vectors. Then, for each query, the system computed cosine similarities between the query vector and all master vectors, selecting the 200 candidates with the highest scores. This process is much faster than fuzzy matching because vector multiplication can be performed in parallel, unlike fuzzy matching, which requires sequential comparison of strings. The selection of TF-IDF character n-grams as the initial filtering method was based on the need for speed, prioritizing high precision. The goal is to filter candidates that are "similar enough" for further processing, rather than immediately selecting the best ones.

With a much smaller search range, more complex and computationally intensive ensemble scoring methods such as BM25 and fuzzy string matching can be run efficiently in subsequent stages without wasting significant time. The choice of a threshold of 200 candidates was pragmatic, balancing coverage and efficiency. The larger the number of candidates, the higher the likelihood of filtering in the correct answer, but the greater the computational burden on the ensemble scoring stage. The value of 200 was chosen because initial testing showed it could maintain adequate coverage, achieving a Recall@200 of 98.36% across the entire dataset, while keeping the overall process within a reasonable time.

The two-stage retrieval approach used in this study, which involves initial TF-IDF-based filtering before proceeding to more complex scoring methods, aligns with patterns that have proven effective in the information retrieval literature. Maryamah et al. [9] in a study of Indonesian drug data retrieval used MPNet and BM25L without an explicit candidate filtering stage, so that all master data was processed directly in each query. The approach proposed in this study improves on this strategy by explicitly separating the retrieval and scoring stages, so that more computationally expensive methods such as fuzzy string matching are only applied to a selected subset of candidates. This strategy aligns with the cascade ranking principle recommended in a survey of modern text-matching techniques [23], where fast methods are used to narrow the search space before more precise methods are applied to selected candidates.

From a clinical implementation perspective, the computational efficiency achieved by this candidate retrieval stage has direct implications for the feasibility of system deployment in healthcare facilities with limited infrastructure. By reducing the search space from 55,078 entries to 200 candidates per query, the system

can process a single mapping request in sufficient time for real-time use without requiring high-powered hardware. This is important in the context of primary healthcare facilities in Indonesia, which generally lack access to GPU clusters or high-powered computing servers. Furthermore, the pragmatically chosen threshold of 200 candidates in this study can serve as a tunable starting point based on the scale of the master data and the computing capacity of each facility, providing flexibility for system adaptation to various deployment contexts.

3.3. Multi-Method Ensemble Results

The multi-method ensemble method used in this study yielded a Top-1 Accuracy score of 68.95% across the entire test data. This result represents a 7.95 percentage point improvement over the initial benchmark score using Euclidean Distance (61.00%) and a 27.95 percentage point improvement over Cosine Similarity (41.00%). The complete metric evaluation results are presented in Table 8.

The significant improvement over both benchmarks is due to the characteristics of each ensemble component. The TF-IDF approach used in this study is character-level, unlike conventional TF-IDF implementations that operate at the word level. Character-based approaches are more robust to spelling variations and typos because a single character change only affects a small portion of the n-grams of the entire text representation. This advantage aligns with the principles proposed in a text-matching survey by [23], which showed that character-based methods are generally more robust to spelling variations than word-based methods in domain-specific text matching tasks. In addition, the combination of five methods in this ensemble scheme allows the weaknesses of one method to be compensated for by the strengths of another. This method has been proven effective across various research areas in information retrieval, where the ensemble approach has been shown to increase the F1-score by up to 3.88% compared to the best single model [31]. The detailed evaluation results are presented in Table 8.

Table 8. Multi-Method Ensemble Results

Evaluation Metrics	Value
Top-1 Accuracy	68.95%
Top-2 Accuracy	78.92%
MRR	0.7393

A significant difference occurred between the Top-2 Accuracy (78.92%) and the Top-1 Accuracy (68.95%), amounting to 9.97%. This gap is noteworthy because it indicates that there were several cases in which the correct answer ranked second, and the system failed to place it in the first rank. This situation typically occurs when two candidates have very similar ensemble scores, leading to ambiguity in determining the final ranking. In this medical study, the Top-2 Accuracy figure of 78.92% indicates that the system has the potential to provide pharmacy staff with two candidate drug prescriptions for confirmation. This approach has proven effective across various AI-based systems in the healthcare sector, significantly reducing manual workload while maintaining expert staff as the primary validators [32].

Based on the evaluation results of this study, the MRR value of 0.7393 aligns with the Top-1 and Top-2 accuracy results. An MRR value close to 1.0 indicates that the system consistently places the correct answer in the top position [33]. An MRR of 0.7393 indicates that the model consistently places the correct answer near the first position. This means that, overall, most correct predictions are in the first position, with only a small percentage in the second. The system rarely places the correct answer outside these two positions. Sample matching results are presented in Table 9.

Table 9. Sample Matching Results

No	Input (free-text)	Top-1 Prediction	TF-IDF	BM25	Fuzzy	Ensemble	Confidence	Top-1	Descriptions
1	(BPJS) GG 100 MG TAB @100 TRIFA	GUAIFENESIN TRIFA 100 MG BOX 100 TABLET	0.991	0.991	0.610	0.707	HIGH	✓	GG=guaifenesin; @100=box 100
2	(BPJS) HYDROCORTISONE 2.5% CREAM 5 GR PHAROS INDONESIA	HYDROCORTISONE ACETATE PHAPROS 2.5% CREAM 5 GRAM TUBE	0.734	1.000	0.743	0.648	MEDIUM	✓	PHAROS=phapros; brand name is different from master
3	(BPJS) ISDN 5 MG TAB	ISOSORBIDE YARINDO	0.547	1.000	0.602	0.589	MEDIUM	✓	ISDN=isosorbide; low TF-IDF due to

No	Input (free-text)	Top-1 Prediction	TF-IDF	BM25	Fuzzy	Ensemble	Confidence	Top-1	Descriptions
	@100 YARINDO	5 MG BOX 100 TABLET							differences in character length
4	(BPJS) METFORMIN 500 MG TAB @200 HEXPHARM	METFORMIN HEXPHARM 500 MG TABLET	0.953	0.976	0.711	0.734	HIGH	✗	HIGH but wrong, packaging information @200 is not recognized by the system
5	ACETYLSISTEIN 200MG OGB DEXA 200MG CAPSUL	ACETYLCYSTEINE OGB DEXA MEDICA 200 MG KAPSUL	0.501	0.492	0.659	0.479	LOW	✓	Severe spelling variations; low scores indicate the need for manual review.

According to the result on Table 9, accuracy was evaluated separately by confidence level. Of the 1,513 predictions categorized as HIGH (score ≥ 0.70), the system achieved Top-1 Accuracy of 87.44%. This value indicates that predictions with high scores are reliable enough to be accepted automatically without manual verification. The MEDIUM category ($0.50 \leq$ score < 0.70) accounted for the largest proportion of predictions, at 77.67% (n = 7,244), with an accuracy of 68.28%. This value is appropriate for a semi-automated scenario in which pharmacist confirmation is still available. Meanwhile, predictions categorized as LOW (score < 0.50) achieved only 28.30% accuracy (n = 569), confirming that low-scoring outputs should not be accepted without manual review. These results indicate that the confidence-scoring mechanism successfully distinguishes predictions that are worthy of full automation from those that require human intervention.

The confidence-based evaluation revealed two distinct failure patterns, both in terms of their nature and clinical impact. In the LOW category, the low accuracy (28.30%) reflects mechanisms operating as designed: as in the case of ACETYLSISTEIN in Table 9, extreme spelling variations resulted in a score of 0.479, which triggered a flag for manual review. However, a more critical failure pattern emerged in the HIGH category, with an accuracy of 87.44%, in which approximately 1 in 8 high-confidence predictions still misreported the drug name. The METFORMIN case in Table 9 illustrates this risk: a score of 0.734 resulted in a HIGH label, allowing the prediction to pass without review, even though the system failed to map the "@200" notation to the "BOX 200" format in the master data due to the limited scope of the normalization dictionary. This type of error is potentially more clinically harmful than the LOW category failures, as there is no mechanism to prevent it from being checked. These findings indicate that the upper limit of HIGH category accuracy is determined more by the completeness of the normalization dictionary than by the quality of the similarity model itself, making dictionary expansion the most impactful development direction for the next iteration of the system.

3.4. Method Comparison Evaluation Results

The graph in Figure 2 compares the Top-1 Accuracy of all methods evaluated in this study. The reference method using Euclidean Distance achieved an accuracy of 61.00%, while Cosine Similarity produced a lower accuracy of 41.00%. The 20 percentage-point difference between the two indicates that the method for measuring vector similarity has a significant impact, even though both use the same vector representation. The low accuracy of Cosine Similarity is likely due to the highly variable string lengths of drug names. Meanwhile, these matching algorithms measure angles between vectors without accounting for differences in document length, limiting their ability to distinguish between drug names with similar keywords. These two baselines are also not designed to handle spelling variations, typos, and word order differences, which are major challenges in free-text prescription data.

The application of Ensemble Scoring, which combines TF-IDF Cosine Similarity with character n-grams, BM25, Token Sort Ratio, Partial Ratio, and Token Set Ratio, resulted in a Top-1 Accuracy of 67.61%. This result represents a 6.61% improvement over the best baseline. This improvement was achieved because the five methods worked in parallel and complemented each other, offsetting the weaknesses of one method by the strengths of the others. The addition of pharmaceutical-domain features in subsequent evaluations increased accuracy to 68.95%. Although this 1.34% increase is relatively small, it can be explained by the limited availability of domain information in the input data: only 2.3% of the data contained explicit packaging information that could be used by the packaging bonus feature. In the subset of data with complete domain information, these features play a crucial role in distinguishing drug names that are generically identical but differ clinically, such as by dosage or manufacturer.

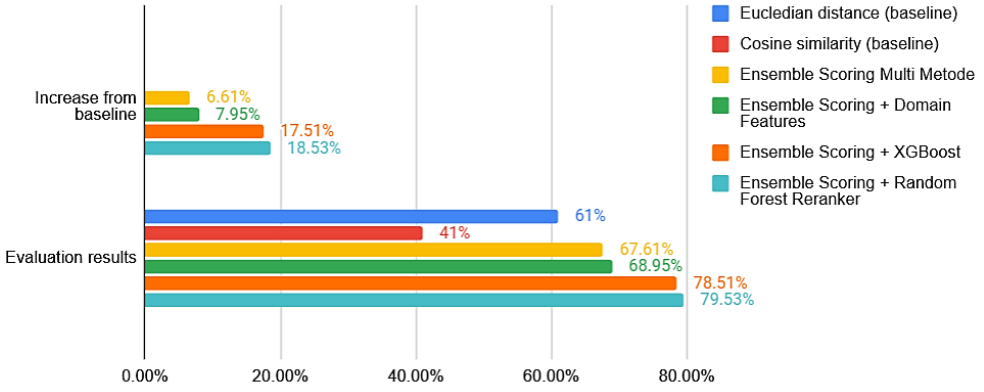


Figure 2. Comparison of the Baseline Method and Final Ensemble Performance

The greatest improvement occurred when a machine learning reranker was applied to the ensemble output. Random Forest achieved a Top-1 Accuracy of 79.53%, and XGBoost achieved 78.51%. Both represent a jump of more than 10% points over other ensemble features. This significant improvement indicates that there are patterns that cannot be captured by a simple linear weighting formula but are successfully learned by the Random Forest model through supervised learning [34]. Random Forest was chosen as the final configuration because it produced the highest accuracy across all evaluations. In addition, RF also provides an interpretive feature importance analysis. Overall, the proposed model achieved a total improvement of 18.53 percentage points over the baseline Euclidean Distance. These results confirm that the stepwise approach from ensemble to machine-learning-based reranker is an effective strategy for mapping drug names from free text to standardized master data.

In practice, the successful integration of the weighted ensemble model with the ML reranker provides concrete recommendations for the architecture of hospital information systems in Indonesia. Through the confidence scoring mechanism, clinical workflows can be transformed into a Human-in-the-Loop system. Pharmacists no longer need to review all prescriptions manually; they can focus on validating data with a low confidence level (LOW Confidence), which has an accuracy of 28.30%, while HIGH prescriptions can be directly mapped automatically to the SATUSEHAT platform. This integration demonstrates that efficient medical data handling does not always require high-spec computer infrastructure, but rather precision in combining feature arrangement algorithms.

4. CONCLUSION

This study successfully implemented an automatic drug-name mapping system for free-text prescriptions, using a multi-method ensemble approach with Random Forest as a reranker. The system achieved a Top-1 Accuracy of 79.53%, representing an improvement of 18.53 percentage points over the baseline Euclidean Distance (61.00%). Feature importance analysis showed that the score difference between the first and second candidates (Top-1 and Top-2 gap) was the strongest predictor of prediction accuracy, contributing 37.21% and indicating that the system's confidence margin was the most informative signal for the reranker. The confidence scoring mechanism proved effective in stratifying prediction reliability: the HIGH category (score ≥ 0.70) achieved 87.44% accuracy and is suitable for automated acceptance; the MEDIUM category ($0.50 \leq$ score < 0.70) achieved 68.28% accuracy and is suitable for semi-automated scenarios with pharmacist confirmation; while the LOW category (score < 0.50) achieved only 28.30% accuracy and always requires manual verification. The dominant MEDIUM proportion (77.67% of the total data) indicates that most cases fall into the gray area requiring human assistance. Therefore, substantial reductions in pharmacist workload can only be achieved if the accuracy of this category is further improved. It is further noted that the candidate retrieval stage achieved a Recall@200 of 98.36%, confirming that the remaining accuracy gap is attributable to ranking errors rather than retrieval failures, and that future improvement efforts should prioritize the reranking component.

The main limitation of this study is the incomplete coverage of the normalization dictionary, as reflected in system failures for spelling variations not covered by the dictionary, including several HIGH predictions that passed without verification but proved incorrect. The addition of the pharmaceutical domain feature resulted in only a 1.34% improvement in accuracy, driven by the low availability of domain information in the input data: only 2.3% of prescriptions explicitly included packaging information in a format recognizable by the system. While this feature has proven crucial for distinguishing entries with identical generic and manufacturer names, its global impact is currently limited by inconsistent input habits among healthcare professionals. It should also be noted that the dataset used comes from a single commercial electronic medical record platform in Indonesia, so the observed naming and abbreviation patterns likely reflect platform-specific

conventions. Validation on a more diverse dataset, multi-institutional datasets, particularly from tertiary hospitals with more complex pharmaceutical terminology, is highly recommended. Further development can focus on expanding the normalization dictionary to cover a broader range of brand-to-generic mappings and platform-specific abbreviation patterns. Leveraging transformer-based language models, such as BioBERT or domain-adapted IndoBERT, may also achieve richer semantic text representations that better bridge the gap between brand names and their generic equivalents in free-text prescriptions. Additionally, exploring strategies to improve prediction accuracy in the MEDIUM confidence category, the dominant data group at 77.67%, remains a critical priority for meaningful reduction in pharmacist workload.

REFERENCES

[1] Kementerian Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022 tentang Rekam Medis.* 2022, pp. 1–20.

[2] K. Ulum, I. L. Hilmi, and S. Salman, "Review Artikel : Implementasi dan Evaluasi Peresepan Elektronik Dalam Upaya Menurunkan Kesalahan Pengobatan," *Journal of Pharmaceutical and Sciences*, vol. 6, no. 1, pp. 192–198, Jan. 2023, doi: 10.36490/journal-jps.com.v6i1.19.

[3] W. H. Organization, *Global Burden of Preventable Medication-Related Harm in Health Care : A Systematic Review*. World Health Organization, 2024.

[4] L. Naseralallah, D. Stewart, M. Price, and V. Paudyal, "Prevalence, Contributing Factors, and Interventions to Reduce Medication Errors in Outpatient and Ambulatory Settings: A Systematic Review," *Int. J. Clin. Pharm.*, vol. 45, no. 6, pp. 1359–1377, Dec. 2023, doi: 10.1007/s11096-023-01626-5.

[5] A. E. Amalia and M. Basabih, "Overview of Medication Error Incidence in Hospitals in Various Countries: Literature Review," *Jurnal Administrasi Kesehatan Indonesia*, vol. 11, no. 1, pp. 145–153, May 2023, doi: 10.20473/jaki.v11i1.2023.145-153.

[6] S. Suwarti, E. Rachmani, and E. Rimawati, "Electronic Medical Records in Increasing User Satisfaction: Literature Review," *Management Analysis Journal*, vol. 14, no. 1, pp. 14–22, 2025, [Online]. Available: http://maj.unnes.ac.id

[7] A. A'aqoulah *et al.*, "Trends and Types of Medical Errors Reporting Following the Implementation of Electronic Health Record System," 2026, doi: 10.2147/DHPS.S558973.

[8] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) dalam Temu Kembali Informasi pada Dokumen Teks," *SINTESIA: Jurnal Sistem dan Teknologi Informasi Indonesia*, vol. 01, no. 2, pp. 81–88, 2022.

[9] M. Maryamah, G. Wilsen, C. T. Suhalim, R. Septiana, A. Fajar, and M. I. Solihin, "Hybrid Information Retrieval with Masked and Permuted Language Modeling (MPNet) and BM25L for Indonesian Drug Data Retrieval," in *2024 16th International Conference on Knowledge and Smart Technology (KST)*, IEEE, Feb. 2024, pp. 242–247. doi: 10.1109/KST61284.2024.10499674.

[10] M. L. Wallace *et al.*, "Use and Misuse of Random Forest Variable Importance Metrics in Medicine: Demonstrations Through Incident Stroke Prediction," *BMC Med. Res. Methodol.*, vol. 23, no. 1, p. 144, Jun. 2023, doi: 10.1186/s12874-023-01965-x.

[11] H. Le *et al.*, "Rxnorm for Drug Name Normalization: A Case Study of Prescription Opioids in The FDA Adverse Events Reporting System," *Frontiers in Bioinformatics*, vol. 3, Jan. 2024, doi: 10.3389/fbinf.2023.1328613.

[12] I. Azam, K. Jiang, and G. B. Bernard, "Normalizing Health Concepts with Biomedical Embedding and LLMs," 2026. [Online]. Available: https://github.com/Medical-NLP/COVID-19-Sig

[13] Z. Xu, Z. Huang, S. Zhuang, and V. Srikumar, "Distillation versus Contrastive Learning: How to Train Your Rerankers." [Online]. Available: https://github.com/zhichaoxu-shufe/

[14] F. Neha, D. Bhati, and D. K. Shukla, "Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review," *AI*, vol. 6, no. 9, pp. 1–29, Sep. 2025, doi: 10.3390/ai6090226.

[15] M. Triartama Manurung, G. Ngurah, L. Wijayakusuma, I. Putu, and W. Gautama, "Named Entity Recognition for Medical Records of Heart Failure Using a Pre-trained BERT Model," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[16] M. H. Rifki, Y. R. W. Utami, and P. Harsadi, "Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algoritma Naïve Bayes," *JuSiTik : Jurnal Sistem dan Teknologi Informasi Komunikasi*, vol. 7, no. 2, pp. 77–86, Jul. 2024, doi: 10.32524/jusitik.v7i2.1168.

[17] A. Hermawan, I. Jowensen, J. Junaedi, and Edy, "Implementasi Text-Mining untuk Analisis Sentimen pada Twitter dengan Algoritma Support Vector Machine," *JST (Jurnal Sains dan Teknologi)*, vol. 12, no. 1, pp. 129–137, Apr. 2023, doi: 10.23887/jstundiksha.v12i1.52358.

[18] H. Aditya, Ardiansyah, Sidik, and W. Gata, "Analisis Sentimen Penggunaan Twitter Terhadap Penggunaan Cairan Desinfektan Menggunakan Metode Term Frequency-Inverse Document Frequency Dan Support Vector Machine," *Jurnal Ilmu-ilmu Informatika dan Manajemen*, vol. 14, no. 2, pp. 167–

174, 2020, Accessed: May 18, 2025. [Online]. Available: https://media.neliti.com/media/publications/414201-none-8fa7a6de.pdf
[19] Sintia, S. Defit, and G. Widi Nurcahyo, "Product Codefication Accuracy with Cosine Similarity and Weighted Term Frequency and Inverse Document Frequency (TF-IDF)," *Journal of Applied Engineering and Technological Science*, vol. 2, no. 2, pp. 14–21, 2021, Accessed: May 18, 2025. [Online]. Available: https://journal.yrpipku.com/index.php/jaets/article/view/210/175
[20] I. S. Wibowo, A. Witanti, and I. Susilawati, "Keyword Extraction Judul Berita Online Di Indonesia Menggunakan Metode TF-IDF," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 11, no. 1, pp. 99–111, [Online]. Available: http://jurnal.mdp.ac.id
[21] J. Pokrywka, "Passage Retrieval of Polish Texts Using OKAPI BM25 and an Ensemble of Cross Encoders," Sep. 2023, pp. 1265–1269. doi: 10.15439/2023F9253.
[22] P. Jiang and X. Cai, "A Survey of Text-Matching Techniques," *Information (Switzerland)*, vol. 15, no. 6, pp. 1–53, Jun. 2024, doi: 10.3390/info15060332.
[23] S. Pawestri and Y. Suyanto, "Analisis Perbandingan Metode Similarity untuk Kemiripan Dokumen Bahasa Indonesia pada Deteksi Kemiripan Teks Bahasa Indonesia," *Jurnal Media Informatika Budidarma*, vol. 8, no. 3, pp. 1440–1450, Jul. 2024, doi: 10.30865/mib.v8i3.7648.
[24] A. Dzaky, H. Musta'in, A. Sanjaya, and A. B. Setiawan, "Penerapan Regular Expression dan Cosine Similarity pada Uji Kemiripan Kalimat Bahasa Indonesia," *INOTEK*, vol. 9, pp. 154–162, 2025.
[25] Q. Hu, Y. Chen, D. Zou, Z. He, and T. Xu, "Predicting Adverse Drug Event Using Machine Learning Based on Electronic Health Records: A Systematic Review and Meta-Analysis," *Front. Pharmacol.*, vol. 15, pp. 1–17, Nov. 2024, doi: 10.3389/fphar.2024.1497397.
[26] M. M. Alvanof, Bustami, and R. K. Dinata, "Penerapan Algoritma Random Forest dalam Deteksi dan Klasifikasi Ransomware," *Jurnal Elektronika dan Teknologi Informasi*, vol. 5, no. 2, pp. 23–31, 2024.
[27] M. Li *et al.*, "Using Semantic Text Similarity calculation for question matching in a rheumatoid arthritis question-answering system," *Quant. Imaging Med. Surg.*, vol. 13, no. 4, pp. 2183–2196, Apr. 2023, doi: 10.21037/QIMS-22-749/COIF.
[28] O. Rainio, J. Teuho, and R. Klén, "Evaluation Metrics and Statistical Tests for Machine Learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, Dec. 2024, doi: 10.1038/s41598-024-56706-x.
[29] C. T. Hoyt, M. Berrendorf, M. Galkin, V. Tresp, and B. M. Gyori, "A Unified Framework for Rank-based Evaluation Metrics for Link Prediction in Knowledge Graphs," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2203.07544
[30] A. Wang, C. Liu, J. Yang, and C. Weng, "Fine-Tuning Large Language Models for Rare Disease Concept Normalization," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 2076–2083, Sep. 2024, doi: 10.1093/jamia/ocae133.
[31] M. Li *et al.*, "Using Semantic Text Similarity calculation for question matching in a rheumatoid arthritis question-answering system," *Quant. Imaging Med. Surg.*, vol. 13, no. 4, pp. 2183–2196, Apr. 2023, doi: 10.21037/qims-22-749.
[32] L. Liu *et al.*, "Web-Based Application Based on Human-in-the-Loop Deep Learning for Deidentifying Free-Text Data in Electronic Medical Records: Development and Usability Study," *Interact. J. Med. Res.*, vol. 12, pp. 1–17, Aug. 2023, doi: 10.2196/46322.
[33] E. M. Voorhees, "The TREC-8 Question Answering Track Report," 2000.
[34] L. Breiman, "Random Forests," 2001.