ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/373511772

Predicting Quality Medical Drug Data Towards Meaningful Data using Machine Learning

Article *in* International Journal of Advanced Computer Science and Applications · August 2023

DOI: 10.14569/IJACSA.2023.01408114

CITATIONS
0

READS
200

1 author:

Wael AL-zyadat
Al-Zaytoonah University of Jordan
59 PUBLICATIONS **231** CITATIONS

SEE PROFILE

All content following this page was uploaded by Wael AL-zyadat on 31 August 2023.

The user has requested enhancement of the downloaded file.

Predicting Quality Medical Drug Data Towards Meaningful Data using Machine Learning

Suleyman Al-Showarah[1], Abubaker Al-Taie[2], Hamzeh Eyal Salman[3], Wael Alzyadat[4], Mohannad Alkhalaileh[5]
Software Engineering Department, Faculty of Information Technology, Mutah University, Karak, Jordan[1,3]
Computer Science Department, Faculty of Information Technology, Mutah University, Karak, Jordan[2]
Software Engineering Department, Faculty of Science and IT, Al-Zaytoonah University, Amman, Jordan[4]
College of Education, Humanities and Social Sciences, Al Ain University, Al-Ain City, United Arab Emirates[5]

***Abstract*—This research aims to improve the process of finding alternative drugs by utilizing artificial intelligence algorithms. It is not an easy task for human beings to classify the drugs manually, as this requires much longer time and more effort than doing it using classifiers. The study focuses on predicting high-quality medical drug data by considering ingredients, dosage forms, and strengths as features. Two datasets were generated from the original drug dataset, and four machine learning classifiers were applied to these datasets: Random Forest, Support Vector Machine, Naive Bayes, and Decision Tree. The classification performance was evaluated under three different scenarios, which varied the ratio of the training and test data for both datasets, as follows: (i) 80% (training) and 20% (test dataset), (ii) 70% (training) and 30% (test dataset), and (iii) 50% (training) and 50% (test dataset). The results indicated that the Decision Tree, Naive Bayes, and Random Forest classifiers showed superior performance in terms of classification accuracy, with over 90% accuracy achieved in all scenarios. The results also showed that there was no significant difference between the results of the two datasets. The findings of this study have implications for streamlining the process of identifying alternative drugs.**

***Keywords*—classification; alternative drugs; medical; decision tree; Support Vector Machine; Naive Bayes; Random Forest**

I. Introduction

Computers have brought significant technical improvements that have resulted in the creation of huge amounts of data, particularly in pharmaceutical and healthcare systems. The availability of huge amounts of data has increased the need for data mining techniques to produce useful knowledge [1]. Accurate analyses of medical drug data are required to discover appropriate alternative medication for a patient, which is gaining with increasing data in the health care and biomedical communities [2][7].

Raw drugs data requires a clear description and interpretation for analysis purposes, this is to find the similarities between drugs that have the same properties and then to find the alternative drugs [2]. The drugs data has many different attributes collected from different sources. The heterogeneity of drug sources, and the variation of their types, made it an uneasy task for human beings to classify the drugs manually as this needs much longer time and more effort than doing it using classifiers. One of the biggest problems is that big data processing and analysis are challenging to acquire meaningful data to support an accurate medical drug practice [10]. As a result, automated medicines classifiers can assist pharmacists and clinicians in prescribing an acceptable replacement prescription if the desired drug is unavailable, as long as the alternative drug has the same constituent name [1][29].

Quality medical drug data refers to accurate and reliable information about medications, including their uses, dosage, side effects, interactions, and other important details. This information is used by healthcare professionals, researchers, and patients to make decisions about the use and prescribing of drugs. Quality drug data is essential for ensuring safe and effective medication use and is typically obtained from reputable sources such as the FDA, the WHO, and medical literature [1].

Artificial intelligence (AI) can play a role in the collection, analysis, and dissemination of quality medical drug data. For example, AI algorithms can be exploited to mine large amounts of data from various sources, such as clinical trials, and electronic health records, to identify patterns and generate new insights about drugs [14]. AI can also be used to support drug discovery and development by identifying potential new drug candidates, predicting their efficacy and safety, and optimizing their formulation and delivery. AI models can also be used in drug safety monitoring by analyzing electronic health records, clinical trial data, and spontaneous reports to detect safety signals and identify potential risks associated with drugs [24]. Moreover, AI-based chatbots can be used to provide patients with personalized information about their medications, including dosage instructions, potential side effects, and interactions with other drugs. Overall, AI can help to improve the quality of medical drug data by enabling faster, more accurate, and more complete data analysis, and by providing new ways to access and use this information. Many ongoing AI research projects aim to improve the quality of medical drug data, such as Insilico [28]. Medicine [27], Exscientia [25], DeepChem [17], and Numerate [26] all of these focused on drug discovery and development.

Pharmacies and other medical professionals can benefit from using an alternative drugs model by assisting them in classifying drugs based on their chemical properties. Such a proposed model would remarkably shorten the time to identify alternative medicines if the original is not found [1]. It also helps people who are looking for a drug with a lower price and cannot afford to purchase the original drug. So, the proposed model can help them to identify other drugs alternatives that are more affordable to them at the same time, and such options have similar chemical properties to the original option [2]. This research aims to predict alternative medical drug using AI algorithms. The alternative drug has the same chemical

characteristics as the original drugs using several features such as Ingredients, Dose Forms, and Strengths. This goal can be accomplished by proposing a conceptual approach for classifying structured medical data using Machine Learning (ML) algorithms: Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM). This aim can be accomplished by considering the following objectives: (i) to investigate the performance of the different classifiers mentioned above. (ii) to investigate the quality of raw data in terms of data amount, diversity, and minimization. (iii) to propose a conceptual approach for classifying structured medical data using machine learning.

The main contribution that can be provided by the proposed method is in the following: (i) propose an efficient method that is used to deal with semi-structured data and transform it into a structured format, which is a meaningful format used with the classifiers. (ii) provide a new way that is used to label the input drugs according to their properties based on two rules. Each labeling rule takes into account a specific combination of attributes. (iii) investigate the use of different classifiers, study their effect on the accuracy of predicting the correct drugs as well as compare their performances.

The remainder of this paper is structured as follows: Section II presents different machine learning algorithms and discusses their importance in such a study. The third Section III illustrates the related previous work. In Section IV, the proposed approach, four main steps used to build the prediction model is presented and explained. The experimental results and discussion is discussed in Section V. Lastly, Section VII presents our work's conclusion and future work.

II. Background

Predictions from AI refer to forecasts or estimates made by an artificial intelligence (AI) system. These predictions can be made using a variety of techniques, such as machine learning, deep learning, or natural language processing [7]. The accuracy of these predictions will depend on the quality of the data used to train the AI model, as well as the complexity of the model itself. Some examples of predictions made by AI include stock market forecasting, weather forecasting, image or speech recognition, and many more [11][16]. Furthermore, the quality of the data that has been processed and analyzed by a machine learning model to make predictions or forecasts. The quality of the predictions will depend on the accuracy of the model, as well as the quality of the input data used to train the model [24].

Predicting the efficacy and safety of a medical drug is an important task in the drug development process. Artificial intelligence can be used to help predict the effectiveness of a drug by analyzing large amounts of data from preclinical and clinical trials. This includes things such as genetic data, demographics of the patient, and laboratory results [14]. In addition, artificial intelligence can be used to identify potential side effects and interactions with other drugs. One of the popular methods is using machine learning (ML) models to analyze data from drug trials and electronic health records (EHRs) to identify patterns that may indicate a drug's efficacy or potential side effects. Another approach is to use deep learning models to analyze the chemical structure of the drug and predict its potential interactions with other molecules in the body. It is worth noting that AI-based predictions for medical drugs are still in the early stage, and many pharmaceutical companies are actively researching and developing new methods for using AI in drug development. Under the umbrella of AI in drug scope encompass, two branches of drugs are drug discovery and predicted drug.

Drug discovery is a process for identifying and developing new medications while predicting drug efficacy and safety refers to using artificial intelligence techniques to analyze data to make predictions about how a drug will perform in preclinical and clinical trials [29]. Drug discovery involves identifying potential drug targets, synthesizing and testing new compounds, and conducting preclinical and clinical trials to determine a drug's efficacy and safety. This process can take many years and involve a significant investment of time and resources [5]. On the other hand, using AI to predict drug efficacy and safety involves analyzing data from various sources, such as preclinical trial results, electronic health records, and genetic data, to identify patterns that may indicate a drug's effectiveness or potential side effects. This can be done more quickly and efficiently than traditional methods and can help reduce the cost and time associated with drug development [1] [2]. In summary, drug discovery is the process of identifying and developing new drugs from scratch, while using AI to predict drug efficacy and safety is a way to analyze existing data and make predictions about how a drug will perform in preclinical and clinical trials.

Also, this section provides the necessary background to understand how the following classifiers work: Random Forest (RF), Naive Beyes (NB), Decision Tree (DT), and Support Vector Machine (SVM).

Random Forest. Random forest is a learning algorithm for classification and regression tasks. It works by building multiple decision trees at training time, which is the cornerstone for the classification or discrimination regression processes. These multiple decision trees use RF to ensure accurate and reliable prediction [21].

Naive Bayes. Naive Bayes is one of the most famous machine learning algorithms, data analysis, and classification. Specifically, it can be characterized by rapid processing and efficiency in forecasting processes. This classifier is based on the statistical concept, Bayes' theorem. It computes the probability of a given result by verifying what is available and known as Naive because it adheres to the independence assumptions principle. As a result, the relationships between all attributes and features are thought to be independent of one another [18]. So that the Naive Bayes model is trained with the data and its characteristics available in the databases. The model then determines the type of new records and classifies them based on the data and statistics available to it. The formula for Naive Bayes is [18]:

$$P(C|X) = \frac{(P(X|C)P(C))}{(P(X))} \tag{1}$$

Decision Tree. A decision tree is a supervised learning algorithm that continuously divides data according to a specific parameter. As it is a tree that looks like a flow chart that

contains a node called the root and has no edges, while all other nodes have edges and are called leaves (also known as decision nodes) [22][3].

Support Vector Machine. Finding a hyperplane that categorizes the data points in N-dimensional space (N: the number of features) is the goal of the support vector machine algorithm (SVM) [9]. After constructing the hyperplanes, the SVM determines the boundaries between the input classes and the input elements [6].

III. Related Work

There is ongoing research in using AI to predict drug efficacy and safety. Here are the most recent and relevant related works in this field.

One approach is using machine learning (ML) models to analyze data from drug trials and electronic health records (EHR) to identify patterns that may indicate the efficacy or potential side effects of a drug. Nature Medicine used an ML model to analyze data from a clinical trial of a drug to treat Alzheimer's disease and was able to predict which patients would respond well to the drug with high precision [5][20].

Another approach is to use deep learning models to analyze the chemical structure of a drug and predict its potential interactions with other molecules in the body [31]. These models can be trained on large datasets of chemical compounds and their known interactions with proteins, enzymes, and other molecules, to predict potential interactions of new compounds. A deep learning model called a graph convolutional neural network (GCNN) was trained on a dataset of known drug-protein interactions and then used to predict potential interactions of new compounds with a protein called cytochrome P450 3A4 (CYP3A4). The results showed that the model was able to predict potential interactions with high accuracy [19]. Similar research purposes are in [32], a deep learning model called a graph attention network (GAT) was trained on a dataset of known drug-protein interactions and then used to predict potential drug-protein interactions. The results showed that the model was able to predict potential interactions with high accuracy and outperformed traditional machine learning methods.

In [2], Alzyadat et al. proposed an approach to predict a targeted drug for the variety of large data structures measured by a stability scale in the preprocessing phase. Their approach performs quality data analysis using correlation methods to identify feature choices related to mapping data, which concerns the basic methods for predicting data based on the K-mean cluster and decision tree. The result of the prediction of the target drug was used as a principal component analysis (PCA) by distance value.

In [1], Al-Hgaish et al. proposed an approach based on the K-Mean algorithm to maintain the quality of medical drug data toward meaningful data in the data lake by clustering big data scope. The K-Mean clustering is used to form different clusters. Each cluster that was produced represents an alternative drug that is compatible with data lake components. The results show that the approach presented in their paper has achieved 92.7% accuracy.

In [12], Huang et al. proposed an approach to classify unknown drugs and provide assistance for drug screening during the development process. They collected a drug dataset using a web crawler. Based on this dataset, the authors derived an equation to calculate the similarity between drugs and defined similarity calculation equation parameters from a subset of the data. Drug data was categorized using the KNN (K closest neighbor) classifier based on drug similarities. The findings demonstrated that the suggested drug classification model can achieve a 77.7% accuracy value.

In [13], using the DrugBank dataset, Ibrahim et al. suggested a similarity-based machine learning system named "SMDIP". To describe the sparse feature space, they computed drug-drug similarities using an evaluation metric for the available biological and structural information on DrugBank. The chosen DDI (Drug-Drug Interaction) key features are subjected to the deployment of six different ML model types. With the following results: Precision 82%, Recall 62%, F-measure 78%, and Accuracy 79%, SMDIP has demonstrated favorable prediction performance when compared to relevant studies.

In [8], Dang et al. used data consisting of approved drugs of histamine antagonists that are connected to 26,344 Drug-Drug Interactions (DDI) pairs from the DrugBank database. Several classifiers such as Random Forest, Naive Bayes, Logistic Regression, Decision Tree, and XGBoost were used with five-fold cross-validation to approach a large-scale DDIs prediction among histamine antagonist drugs. According to the prediction performance, their model performed better than previously published works on DDI prediction with the best Precision of 78.8%, Recall of 92.1%, and F1-score of 83.8% among 19 given DDIs types.

The differences between our work with other studies are as follows:

1- A new methodology not used before in the previous studies to achieve the aim of this study by using the two datasets produced from the original dataset.

2- To the best of our knowledge, we have not come across any study conducted by using the four classifiers that are used in this study, especially, on this dataset (i.e. FDA) in order to achieve the aim.

IV. The Proposed Approach: Building Prediction Model

The proposed approach is presented in this section. In the beginning, we give a holistic view of the approach. The approach's steps are then detailed in subsequent subsections.

A. Overview

The essential steps of the proposed medicine categorization approach are explained in this section. Fig. 1 shows the four steps for the suggested model after importing the dataset to obtain an alternative drug. These steps are as follows: 1) pre-processing, 2) feature extraction, 3) applying heuristic-based rules, and 4) applying different classifiers with different scenarios The following is an explanation of the proposed approach, which consists of the following steps:

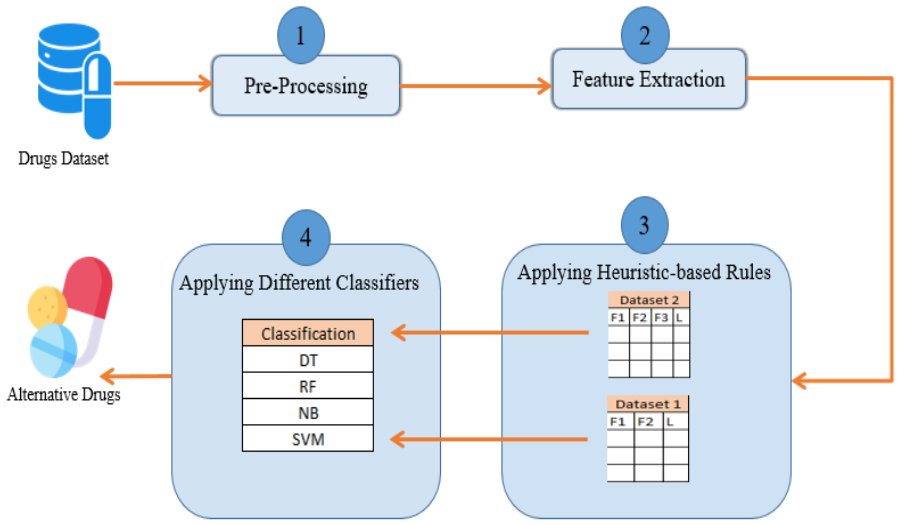

Fig. 1. An overview of the proposed approach (F: Feature, L: Label).

TABLE I. SELECTED FEATURES FOR TRAINING

Features	Type	Description
Ingredients	textual	The active ingredients of the medicine
Strengths	numerical	Active component effect
Dose Forms	numerical	The medicine dose route or form

B. Step 1: Pre-processing

This is the first step in the proposed approach, as shown in Fig. 1. Pre-processing is the step-in machine learning and natural language processing where raw data is cleaned, transformed, and organized in a format that is suitable for the model to train. This can include tasks such as tokenization, stemming, and removing stop words, as well as more complex tasks such as creating new features or dealing with missing data. Pre-processing is a crucial step in building a successful machine learning model, as the quality and structure of the input data can have a big impact on the performance of the model. In this step, you consider the content of the data (indexing) through two aspects. The first aspect performs a vertical process that indexes each attribute with missing data, incomplete data, outliers, or empty data (anything that is zero or none). The second aspect of data pre-processing is horizontal, where duplication of all rows is considered. The importance of this step is to avoid ambiguity and to increase the meaningful data set. On the other hand, the relationship between these attributes, which are best derived by feature selection will be clear [2].

C. Step 2: Features Selection

The important attributes were extracted in this step based on the previous studies and opinions of specialists of pharmacists and doctors. The database features that were picked for the investigation are displayed in Table I.

D. Step 3: Applying Heuristic-based Rules

Heuristic-based rules are a type of rule-based method used in artificial intelligence and natural language processing. These rules are based on heuristics, which are general problem-solving strategies or "rules of thumb" that are used to make decisions or solve problems. Heuristic-based rules use these rules of thumb to make decisions about how to process or understand natural language input. Heuristic-based rules are generally simple and easy to understand, but they can be prone to errors and biases. They are useful in situations where the data is well understood, and the rules can be defined to cover most cases. However, they may not be able to handle more complex or ambiguous input. An example of using heuristic-based rules in medical drug data would be to identify drug interactions based on a set of predefined rules. For example, a rule may state that if a patient is taking drug *A* and drug *B*, there is a high risk of interaction and the dosage of one or both drugs should be adjusted. Another example is extracting information from electronic medical records (EMR) using heuristic-based rules. The rules can be defined to identify specific patterns in the text, such as identifying the name of a drug, the dosage, the frequency of administration, and the duration of treatment. Once these patterns are identified, the information can be extracted and organized into a structured format for analysis. Heuristic-based rules are also used to classify clinical notes from EMR, for example, a rule can be defined that states if a patient is complaining of chest pain and shortness of breath, then it is likely a case of Angina [15]. In summary, heuristic-based rules are an efficient and cost-effective way to extract and analyze medical drug data. They are particularly useful when the data is well-defined, and the rules can be easily formulated to cover most cases. However, they may not be as robust as other methods in handling complex or ambiguous input.

In this study, two heuristic-based rules (rule 1 and rule 2) were used in the proposed approach to produce two datasets, as follows.

1) Rule 1: Two Similar Features: Dataset 1 was produced from the original dataset after applying rule 1, which is explained in (Equation 2 below). The content of this rule suggests that all drugs having similar ingredients, and strength values would have the same label number.

$$L1 = sim(D_i(x_1, x_2), D_j(x_1, x_2)) \quad (2)$$

From Equation 2, suppose that there are two drugs, D_i, D_j each drug has two attributes, i.e. the ingredient, and strength represented by the variables x1, x2. As rule 1 suggests that if these two drugs have the same or similar values of x1, x2, it can be said that both drugs have the same category or label (L).

2) Rule 2: Three Similar Features: Dataset 2 was produced using the same idea that was used to produce dataset 1. However, the content of this rule suggests that all drugs having similar ingredient, strength, and dose values would have the same label number, as stated in Equation 3.

$$L2 = sim(D_i(x_1, x_2, x_3), D_j(x_1, x_2, x_3)) \quad (3)$$

The equitation can be read as follows. x1, x2, x3 are ingredient, dose form, and strength, respectively. D_i, D_j are two drug items. Sim is similarity measurement, and L represents the label ID.

E. Step 4: Applying Different Classifiers with Different Scenarios

In this final step, the following classifiers are applied in this study: Decision Tree, Random Forest, Support Vector Machine, and Naive Bayes. These classifiers are applied individually to the datasets, which are obtained in the previous step, according to their description in the background section. The application of these classifiers on each dataset (dataset 1 and dataset 2) is done according to different scenarios. In each scenario, a percentage of the dataset's records is considered to train the prediction model while other records are used to test the prediction model.

V. Experiments and Evaluation

In this section, the dataset used to evaluate the proposed approach is described, and the evaluation procedure and metrics are listed.

A. Dataset Description

The dataset utilized in this investigation is from the Food and Drug Administration (FDA) and may be found in the following link(https://www.fda.gov/drugs/drug-approvals-anddatabases/drugsfda-data-files). This dataset consists of 14 features and 37,071 records. The selected features and the description for each feature of the dataset are based on the specialists' viewpoints as shown in Table I. The dataset was approved by FDA and this took several years and involved multiple stages, including preclinical testing, phases of clinical trials, and a review of the drug's safety and efficacy by an advisory committee [1].

The reliability of a dataset refers to the consistency and accuracy of the data it contains. In the case of datasets such as *clinicaltrials.gov* and the FDA's drug approval dataset, the data is typically considered to be reliable as it is collected and compiled by reputable government agencies with strict oversight and regulations in place. The features and properties of drug approval datasets such as *clinicaltrials.gov* and the FDA's drug approval dataset include [1][2]:

1- Drug Information: these datasets contain information on drugs that are currently in development, have been approved, or have been withdrawn from the approval process. This information includes the drug's name, active ingredients, intended use, and the company developing the drug.

2- Study Information: these datasets also contain information on the clinical trials that have been conducted to evaluate the safety and efficacy of the drugs. This includes the study design, the number of participants, the inclusion and exclusion criteria, and the primary and secondary outcome measures.

3- Status Information: these datasets provide information on the current status of the drug's development and approval. This includes whether the drug is currently in preclinical testing, phase 1, 2, or 3 clinical trials, or has been approved or withdrawn by the FDA.

4- Search and Filter Capabilities: these datasets are searchable and can be filtered by various criteria such as drug name, condition, company, and study status.

5- Publicly Available: these databases are publicly available and can be accessed by anyone with an internet connection.

6- Regularly Updated: the data in these datasets is updated regularly as new information becomes available.

B. Research Questions and Evaluation Metrics

In this study, two research questions are answered. These questions are in the following:

- **RQ1**: To what extent the proposed approach is accurate to suggest alternative drugs? This research question aims to show the ability of the proposed approach to suggest the most suitable alternative drugs.
- **RQ2**: To what extent the proposed approach is comparable to the most recent works in the subject? This research question aims to measure the efficiency of the proposed approach when it is compared to the research works in the literature.

To address the first research question (RQ1), The obtained results are evaluated using well-known measures in this subject [4][23]. These measures are as follows: Precision, Recall, F-measure, and Accuracy. The values of these measures take a range of [0-1]. We looking to have values near to the one in all considered measures. The equations of these measures are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

TABLE II. CLASSIFIERS RESULTS OF SCENARIO 1.

Classifier	Evaluation Measure	Dataset 1	Dataset 2
RF	F1-measure	92.34%	89.73%
	Recall	93.61%	91.62%
	Precision	91.79%	88.64%
	Accuracy	93.61%	90.27%
NB	F1-measure	97.10%	98.48%
	Recall	96.88%	98.73%
	Precision	97.59%	98.36%
	Accuracy	96.88%	98.58%
DT	F1-measure	97.95%	98.47%
	Recall	98.27%	98.73%
	Precision	97.85%	98.34%
	Accuracy	98.27%	98.73%
SVM	F1-measure	2.85%	3.44%
	Recall	9.53%	10.48%
	Precision	1.16%	2.28%
	Accuracy	8.98%	10.48%

TABLE III. CLASSIFIERS RESULTS OF SCENARIO 2.

Classifier	Evaluation Measure	Dataset 1	Dataset 2
RF	F1-measure	91.28%	87.58%
	Recall	92.83%	90.19%
	Precision	90.65%	86.09%
	Accuracy	92.83%	90.19%
NB	F1-measure	95.04%	96.94%
	Recall	95.83%	97.52%
	Precision	94.92%	96.64%
	Accuracy	95.83%	97.52%
DT	F1-measure	96.47%	96.88%
	Recall	97.10%	97.52%
	Precision	96.25%	96.54%
	Accuracy	97.10%	97.52%
SVM	F1-measure	1.45%	2.44%
	Recall	8.56%	9.24%
	Precision	0.16%	1.75%
	Accuracy	7.97%	9.49%

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

where TP, TN, FP, and FN indicate to true positive class, true negative class, false positive, and false negative class, respectively.

The classifiers used in this study(i.e. Random Forest (RF), Decision Tree (DT), Naive Bayesian (NV), and the Support Vector Machine (SVM)) were applied to each dataset separately; dataset 1 and dataset 2. However, in the training and test dataset of each classifier, three scenarios were used in the study, as follows:

1- **Scenario 1:** 80% of the dataset's records are used for training while 20% of dataset's records are used for testing.

2- **Scenario 2:** 70% of the dataset's records are used for training while 30% of dataset's records are used for testing.

3- **Scenario 3:** 50% of the dataset's records are used for training while 50% of dataset's records are used for testing.

To address the second research question (RQ2), a comparison was made between the proposed approach in this study with other recent and relevant works in terms of Precision, Recall, F-measure, and Accuracy metrics.

VI. RESULTS AND DISCUSSION

This section explains the experimental results from testing the model in different scenarios. As this study was designed for three scenarios, each has two datasets based on the two rules. The results of classifiers are represented by several evaluation criteria: classification Accuracy, F-measure, Recall, and Precision.

A. Research Question (RQ1)

1) The Results of Scenario 1: this subsection presents the results of Scenario 1. This scenario was built based on the percentage of the samples of the dataset used in the experiment. 80% of the records were used to train the dataset while 20% of the records were used to test the dataset. This scenario applied to two datasets; the four classifiers were conducted on each dataset. Table II shows the results of Scenario 1. The results show that DT has achieved high classification accuracy for dataset 1 (98.27%), and dataset 2 (98.73%) compared to other classifiers used in the study. The high accuracy of the Decision Tree over the other classifiers can be attributed to the fact that these kinds of algorithms, i.e., the tree-based classifiers, are less prone to overfitting. At the same time, training makes them robust and rigid against outliers and misclassification [30]. Both the RF and the DT outperformed the SVM, as can be seen. Also, we can see that there was no big difference between the results of dataset 1 and dataset 2. This demonstrates that using two attributes can achieve the study's goal of finding an alternative drug with the strongest effect by using two attributes.

2) The Results of Scenario 2: this subsection presents the results of Scenario 2. This scenario was built based on the percentage of the samples of the dataset used in the experiment. 70% of the records were used to train the dataset, while 30% of the records were used to test the dataset. As can be seen from Table III, it can be derived the same observations noticed for scenario 1. The average performance of each one of the NB, DT, and RF was also higher than 90% which indicates that these three classifiers are good enough to handle the input data. As an observation, the SVM classifier still has minor performance outcomes compared to the other classifiers involved in the study. This is for the same interpretation mentioned in Scenario 1. Also, we can see that there was no big difference between the results of dataset 1 and dataset 2. This proves that using two attributes can achieve the aim of this study and find an alternative drug for the strongest effect of choosing the two attributes.

3) The Results of Scenario 3: this subsection presents the results of Scenario 3, As can be seen from Table IV. This

TABLE IV. CLASSIFIERS RESULTS OF SCENARIO 3.

Classifier	Evaluation Measure	Dataset 1	Dataset 2
RF	F1-measure	89.31%	85.60%
	Recall	91.46%	88.77%
	Precision	88.27%	83.84%
	Accuracy	91.46%	88.77%
NB	F-measure	91.64%	92.81%
	Recall	93.26%	94.41%
	Precision	91.00%	91.92%
	Accuracy	93.26%	94.41%
DT	F1-measure	93.29%	92.78%
	Recall	94.77%	94.41%
	Precision	92.54%	91.87%
	Accuracy	94.77%	94.41%
SVM	F1-measure	6.15%	5.59%
	Recall	7.18%	6.99%
	Precision	5.57%	4.86%
	Accuracy	7.18%	6.99%

scenario was built based on the percentage of the records of the dataset used in the experiment. 50% of the records were used to train the dataset while 50% of the records were used to test the dataset. Not far from the results in both Scenario 1 and 2, it can be realized that the experimental results in Scenario 3 for the tested classifiers have almost similar experimental results for the same reasons mentioned in both Scenario 1 and Scenario 2. Also, we can see that there was no big difference between the results of dataset 1 and dataset 2. This proves that using two attributes can achieve the aim of this study and find an alternative drug for the strongest effect of choosing the two attributes.

As a summary, we can note that the DT classifier outperforms other studied classifiers (RF, NV, and SVM) in terms of Precision, Recall, F-measure, and Accuracy. Also, it is noted that SVM always produced unsatisfactory results. This is based on the experimental results shown in Table II, III, and IV.

B. Research Question 2 (RQ2)

The most recent and relevant work in the literature is the work proposed by Dang et al.[8]. They conducted their study on drug data. Also, various classification algorithms were applied in their studies such as Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and XGBoost. They used the Precision, Recall, F-measure to evaluate the obtained results, but we used the Accuracy measure in addition to Precision, Recall, and F-measure to evaluate our proposed model. Al-Hgaish et al.[1] conducted their study on the same as our dataset but using clustering algorithms. Their study is to maintain the quality of medical drug data toward meaningful data in the data lake by clustering big-data scope using K-Mean Algorithm. They were focused only on analyzing the dataset. Huange et al.[12] conducted their study to classify unknown drugs and provide assistance for drug screening during the development process. They collected the drug dataset using a Web crawler and applied the accuracy as a metric using the k-nearest neighbor classifier. In our study, four classifiers were used (Random Forest, Naive Beyes, Decision Tree, and Support Vector Machine). Also, two datasets were analyzed based on applying two heuristic-based rules. In addition, three different scenarios were applied for each dataset using the following metrics in the analysis: Precision, Recall, F1-measure, and Accuracy.

Table V shows the comparison results among the most recent and relevant works (mentioned above) in this subject. It has been noted that from Table V that the proposed model has outperformed three modern methods published in the past few years in terms of accuracy Dang et al.[8], Al-Hgaish et al.[1], and Huange et al.[12]. This is because of the use of a new methodology as well as applying classifiers that have not been used before in previous studies of the dataset (i.e., the Food and Drug Administration).

VII. CONCLUSIONS AND FUTURE WORK

The goal of this study is to predict quality medical drug data toward meaningful data from an input drug dataset. The alternative drug has the same chemical characteristics as the original drugs have several features: ingredients, dose forms, and strengths. This aim can be accomplished by considering the following objectives: (i) to investigate the performance of different classifiers (i.e., Decision Tree, Random Forest, Support Vector Machine, and the Naive Bayesian) on the drugs dataset. (ii) to investigate the quality of raw data in terms of data amount, diversity, and minimization. (iii) to propose a conceptual approach for classifying structured medical data using machine learning. The experiments were conducted on three scenarios for the following classifiers: Decision Trees, Random Forest, Support Vector Machine, and Naive Bayesian. The obtained results indicated that the Decision Tree, Naive Bayes, and Random Forest classifiers showed superior performance in terms of classification accuracy, with over 90% accuracy achieved in all scenarios. The results also showed that there was no significant difference between the results of the two generated datasets. The findings of this study have implications for streamlining the process of identifying alternative drugs. When it comes to the performance of the Support Vector Machine, it can be realized that it has a major degradation in performance.

Future work involves exploring the use of more advanced machine-learning techniques to improve the accuracy and performance of the classifiers. Another avenue for further research would be to include more features and variables in the analysis to provide a more comprehensive evaluation of the drugs. Additionally, it would be beneficial to compare the results of this study with other existing drug classification systems to identify any areas for improvement. Finally, conducting user studies and gathering feedback from medical professionals could provide valuable insights into the real-world applicability of the proposed approach and identify any potential limitations.

REFERENCES

[1] A. Al-Hgaish, W. Alzyadat, M. Al-Fayoumi, A. Alhroob, and A. Thunibat. Preserve quality medical drug data toward meaningful data lake by cluster. *International Journal of Recent Technology and Engineering*, 8:270–277, 2019.

[2] W. Alzyadat, M. Muhairat, A. Alhroob, and T. Rawashdeh. A recruitment big data approach to interplay of the target drugs. *Int. J. Advance Soft Compu. Appl*, 14(1), 2022.

TABLE V. COMPARISON TABLE WITH RELATED WORKS.

	Precision	Recall	F-Measure	Accuracy
Dang et al. (2021)[8]	78.8%	92.1%	83.8%	—
Al-Hgaish et al. (2019)[12]	—	—	—	92.7%
Huang et al. (2017)[1]	—	—	—	77.7%
Our Proposal	97.85%	98.27%	97.95%	98.27%

[3] P. Argentiero, R. Chin, and P. Beaudet. An automated approach to the design of decision tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(1):51–57, 1982.

[4] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., USA, 1999.

[5] D. S. Battina. The role of machine learning in clinical research: Transforming the future of evidence generation. *FUTURE*, 4(12), 2017.

[6] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020.

[7] H. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan. Advancing drug discovery via artificial intelligence. *Trends in pharmacological sciences*, 40(8):592–604, 2019.

[8] L. H. Dang, N. T. Dung, L. X. Quang, L. Q. Hung, N. H. Le, N. T. N. Le, N. T. Diem, N. T. T. Nga, S.-H. Hung, and N. Q. K. Le. Machine learning-based prediction of drug-drug interactions for histamine antagonist using hybrid chemical features. *Cells*, 10(11), 2021.

[9] R. Gandhi. Support vector machine — introduction to machine learning algorithms. 2018.

[10] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. E. Whang. Goods: Organizing google's datasets. SIGMOD '16, New York, NY, USA, 2016. Association for Computing Machinery.

[11] E. Hamadaqa, A. Abadleh, A. Mars, and W. Adi. Highly secured implantable medical devices. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 7–12, 2018.

[12] D. G. Huang, L. Guo, H. Y. Yang, X. P. Wei, and B. Jin. Chemical medicine classification through chemical properties analysis. *IEEE Access*, 5:1618–1623, 2017.

[13] H. Ibrahim, A. M. El Kerdawy, A. Abdo, and A. Sharaf Eldin. Similarity-based machine learning framework for predicting safety signals of adverse drug–drug interactions. *Informatics in Medicine Unlocked*, 26:100699, 2021.

[14] K.-K. Mak and M. R. Pichika. Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*, 24(3):773–780, 2019.

[15] G. G. Marewski JN. Heuristic decision making in medicine. *Dialogues Clin Neurosci. 2012*, 14(1):77–89, 2022.

[16] A. Mars, A. Abadleh, and W. Adi. Operator and manufacturer independent d2d private link for future 5g networks. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2019.

[17] A. J. Minnich, K. McLoughlin, M. Tse, J. Deng, A. Weber, N. Murad, B. D. Madej, B. Ramsundar, T. Rush, S. Calad-Thomson, et al. Ampl: a data-driven modeling pipeline for drug discovery. *Journal of chemical information and modeling*, 60(4):1955–1968, 2020.

[18] T. R. Patil and S. S. Sherekar. Performance analysis of naive bayes and j 48 classification algorithm for data classification. 2013.

[19] M. Qiu, X. Liang, S. Deng, Y. Li, Y. Ke, P. Wang, and H. Mei. A unified gcnn model for predicting cyp450 inhibitors by using graph convolutional neural networks with attention mechanism. *Computers in Biology and Medicine*, 150:106177, 2022.

[20] S. Qiu, M. I. Miller, P. S. Joshi, J. C. Lee, C. Xue, Y. Ni, Y. Wang, D. Anda-Duran, P. H. Hwang, J. A. Cramer, et al. Multimodal deep learning for alzheimer's disease dementia assessment. *Nature communications*, 13(1):1–17, 2022.

[21] M. Ristin-Kaufmann. Large-scale image recognition with random forests. ETH-Zürich, 2015.

[22] S. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991.

[23] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA, 1986.

[24] M. Schauperl and R. A. Denny. Ai-based protein structure prediction in drug discovery: Impacts and challenges. *Journal of Chemical Information and Modeling*, 62(13):3142–3156, 2022.

[25] E. Smalley. Ai-powered drug discovery captures pharma interest. *Nature Biotechnology*, 35(7):604–606, 2017.

[26] N. Stephenson, E. Shane, J. Chase, J. Rowland, D. Ries, N. Justice, J. Zhang, L. Chan, and R. Cao. Survey of machine learning techniques in drug discovery. *Current drug metabolism*, 20(3):185–193, 2019.

[27] N. J. Sucher. The application of chinese medicine to novel drug discovery. *Expert opinion on drug discovery*, 8(1):21–34, 2013.

[28] G. C. Terstappen and A. Reggiani. In silico research in drug discovery. *Trends in pharmacological sciences*, 22(1):23–26, 2001.

[29] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[30] K. Vijayakumar and C. Saravanakumar. *Multilevel Mammogram Image Analysis for Identifying Outliers: Misclassification Using Machine Learning*, pages 161–175. Springer Singapore, Singapore, 2021.

[31] J. You, R. D. McLeod, and P. Hu. Predicting drug-target interaction network using deep learning model. *Computational biology and chemistry*, 80:90–101, 2019.

[32] L. Zangari, R. Interdonato, A. Calió, and A. Tagarelli. Graph convolutional and attention models for entity classification in multilayer networks. *Applied Network Science*, 6(1):1–36, 2021.

View publication stats