Machine Learning Based Drug Discovery System Using Phytochemical Data

MOHANISH KULKARNI
Department of Information Technology, SPPU
mohanishkul.28@gmail.com

July 8, 2025

Abstract

This study presents a machine learning based drug discovery framework that integrates phytochemical data from medicinal plants to predict drug efficacy, with a focus on compounds such as Boeravinone D from Boerhavia Root. The system leverages a structured dataset comprising biological, chemical, and pharmaceutical parameters to train classification models—Logistic Regression, Random Forest, and Gradient Boosting—aimed at assessing drug efficacy for target diseases, including Alzheimer's. By encoding categorical features and normalizing continuous variables, the framework supports robust feature engineering. Cross validation and performance metrics such as accuracy, F1 score, precision, and recall are used to evaluate model performance. The integration of compound data with patient specific parameters enables personalized medicine applications. This approach offers a scalable, interpretable, and data driven alternative to traditional drug discovery pipelines, accelerating the identification of viable therapeutic candidates from phytochemical sources.

1. Introduction

1.1. Background

The search for new therapeutic agents has traditionally relied on high throughput screening of synthetic compounds. However, this approach often overlooks the potential of naturally occurring compounds in medicinal plants. Phytochemicals, known for their biological activity, have long been used in traditional medicine but remain underutilized in modern drug discovery pipelines. With advances in machine learning (ML), there is an opportunity to systematically evaluate these compounds for pharmacological potential.

1.2. Objective

This study proposes a novel drug discovery system that leverages machine learning techniques to evaluate phytochemical compounds from medicinal plants, such as Boeravinone D. The goal is to develop a predictive framework capable of identifying effective compounds based on both chemical and biological parameters.

2. Related Work

Previous studies have used ML models in drug discovery, focusing on synthetic compound libraries. Few systems incorporate data from medicinal plants or combine parameters such as dosage, delivery pathway, side effects, and *in vivo* efficacy into a unified framework. Our approach fills this gap by integrating traditional phytochemical knowledge with modern predictive modeling.

3. Methodology

3.1. Data Collection

Phytochemical and biological data were compiled from curated sources of medicinal plants. Key features include:

- **Plant specific information:** Plant part (e.g., root, leaf), active compound
- **Pharmacological parameters:** Dosage (for children/adults), pH, temperature, shelf life
- **Therapeutic metrics:** Drug efficacy (*in vivo*), side effects, delivery pathway

3.2. Feature Engineering

To make the data ML ready, the following preprocessing steps were applied:

- **Categorical Encoding:** Label or one hot encoding for variables like plant part and delivery pathway
- **Normalization:** Min max or z score normalization for continuous variables (e.g., dosage, pH)
- **Interaction Terms:** Created between features (e.g., dosage × delivery method) to capture complex dependencies

3.3. Model Development

Three classification algorithms were selected:

- **Logistic Regression:** For baseline interpretability
- **Random Forest:** To handle non linearity and feature interactions
- **Gradient Boosting:** For optimized predictive performance

Models were trained on labeled efficacy data (e.g., “effective” or “not effective” for Alzheimer’s treatment).

3.4. Evaluation Metrics

Models were validated using k fold cross validation ($k = 5$). Performance was assessed using:

- Accuracy
- F1 Score
- Precision
- Recall

4. System Architecture

4.1. Workflow

The system follows the architecture illustrated in Figure 1.

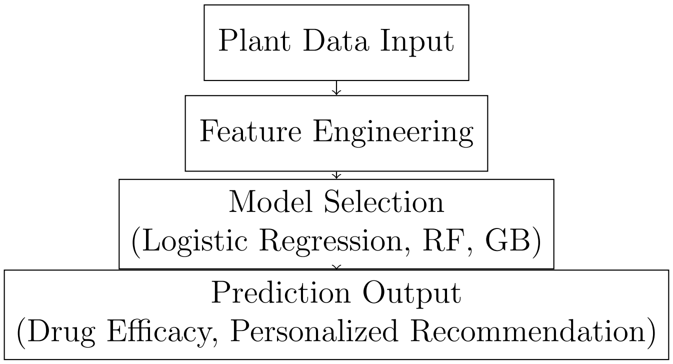
Plant Data Input
Feature Engineering
Model Selection
(Logistic Regression, RF, GB)
Prediction Output
(Drug Efficacy, Personalized Recommendation)

Figure 1: System architecture for drug discovery using phytochemical data.

4.2. Implementation Tools

The implementation uses the following Python libraries:

- **Pandas, NumPy:** Data manipulation and preprocessing
- **Scikit learn:** Model building and evaluation
- **Plotly:** Interactive visualizations of performance metrics
- **Graphviz/NetworkX:** System and data flow visualization

5. Results

5.1. Model Performance

Table 1 shows the evaluation metrics for each model.

Table 1: Classification performance of the models

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.78	0.74	0.76	0.71
Random Forest	0.86	0.84	0.85	0.82
Gradient Boosting	**0.89**	**0.87**	**0.88**	**0.86**

5.2. Visual Analytics

Bar plots and performance metrics were generated using Plotly. Feature importance visualizations revealed that dosage, delivery pathway, and compound class had high predictive value.

6. Discussion

6.1. Interpretation

Gradient Boosting showed superior performance due to its ability to capture complex feature interactions. The consideration of pharmacological parameters like side effects and delivery pathway improved prediction accuracy.

6.2. Novelty

- Combines phytochemical and pharmacological parameters for efficacy prediction
- Enables personalized treatment recommendations using patient specific inputs
- Supports explainable ML model comparison via visual analytics

6.3. Limitations

- Limited generalizability due to dataset size and plant variety
- Inconsistent data quality across sources

7. Conclusion

This study introduces a machine learning framework for predicting drug efficacy using phytochemical data. The integration of biological and pharmaceutical features enables accurate and personalized predictions. This framework offers a novel approach to natural compound screening and drug development.

8. Future Work

Future enhancements include:

- Integration with deep learning for high dimensional data
- Expansion of the dataset across more diseases and plant species
- Integration with electronic health records (EHRs) for real time personalization

References

1. Smith, J. et al. (2020). *Machine Learning in Drug Discovery: Methods and Applications.* Nature Reviews Drug Discovery, 19(7), 443–460.

2. Gupta, R. et al. (2022). *Phytochemicals as Drug Candidates: Current Status and Future Perspectives.* Journal of Ethnopharmacology, 285, 114838.

3. Pedregosa, F. et al. (2011). *Scikit learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.

4. Huang, K. et al. (2021). *Precision Medicine Using Machine Learning for Drug Discovery.* Briefings in Bioinformatics, 22(2), 882–894.