Drug Repurposing Prediction Using Machine Learning Algorithms

Priscilla Yinzime,Michaelangelo Eldridge, Om Patel, Sohini Sahukar
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA
Email: {pyinzime,meldridge1,ssahukar,opatel8}@hawk.illinoistech.edu

***Abstract*—Drug repurposing aims to identify new therapeutic uses for existing drugs, offering a promising alternative to the long, costly, and uncertain process of de novo drug discovery. The increasing availability of biomedical data has motivated the development of machine-learning-based frameworks for uncovering hidden relationships between drugs and diseases. In this work, we develop a complete machine-learning pipeline for drug repurposing prediction using integrated data derived from DrugBank, the Comparative Toxicogenomics Database (CTD), and the BioSNAP Disease–Drug network. After merging these heterogeneous resources, we construct a unified dataset of approximately 1.2 million drug–disease pairs and engineer structural, biological, and relational features. We subsequently evaluate multiple models including LightGBM, XGBoost, CatBoost, and a deep neural network. Although each model captures different types of relationships, XGBoost demonstrates the strongest overall performance, particularly for Top-K ranking. This work highlights the value of machine learning for prioritizing repurposing candidates and underscores the importance of data integration in biomedical prediction tasks.**

***Index Terms*—Drug repurposing, machine learning, biomedical data integration, gradient boosting, deep learning.**

I. Introduction

Drug discovery is traditionally a lengthy and expensive endeavor, often requiring more than a decade of development and over a billion dollars in cost. Many potential drugs fail during clinical trials due to toxicity, lack of efficacy, or other unforeseen issues. Because of this, drug repurposing, identifying new therapeutic applications for existing drugs, has become an increasingly important strategy in modern biomedical research. Drugs that have already been approved for one indication have established safety profiles, substantially reducing the risk, cost, and overall development time required for identifying new uses.

Several historical examples validate the success of drug repurposing. Aspirin, once used solely as an analgesic, was later found to reduce the risk of cardiovascular events. Sildenafil, initially developed for hypertension, was repositioned for erectile dysfunction and, later, pulmonary hypertension. These cases demonstrate the value of identifying new therapeutic roles for existing drugs, especially when underlying biochemical or phenotypic connections are not immediately obvious.

In recent years, machine learning (ML) has shown great promise in biomedical applications, particularly in uncovering complex nonlinear relationships hidden within large datasets. ML-based drug repurposing frameworks typically rely on chemical structures, gene expression signatures, disease ontologies, protein interactions, or network connectivity features. However, integrating multiple datasets remains a challenging task, as each database uses its own identifiers, formats, and assumptions.

This project presents a structured machine-learning pipeline for drug repurposing prediction. We focus on integrating DrugBank, CTD, and BioSNAP to create a unified and consistent drug–disease dataset. We then evaluate four predictive models;LightGBM, XGBoost, CatBoost, and a deep neural network to determine which approaches best capture repurposing-relevant patterns. Our primary evaluation metric is Top-K accuracy, a practical measure for repurposing where the goal is to generate a shortlist of promising disease targets for further investigation.

II. Related Work

Drug repurposing research spans multiple areas including network biology, cheminformatics, computational pharmacology, and machine learning. Early computational approaches relied heavily on chemical similarity: drugs with similar molecular structures or physicochemical properties were assumed to have similar biological effects. Other methods used gene expression signatures to compare drug-induced cellular activity with disease-associated expression patterns.

Network-based approaches represent one of the leading techniques in modern computational repurposing. In these models, drugs, diseases, proteins, and pathways form a multilayer heterogeneous network. Algorithms such as random walks, network propagation, and link prediction have been used to identify novel drug–disease associations based on network connectivity patterns.

Machine-learning-based strategies introduce flexibility by allowing models to learn predictive features rather than relying on predefined similarity metrics. Gradient boosting methods, especially XGBoost and LightGBM, have become popular in biomedical data science due to their robustness, interpretability, and scalability. Deep learning has also gained attention through graph convolutional networks, sequence models, and autoencoders. However, deep learning models typically require highly structured networks or large-scale omics datasets.

Our work follows a structured ML approach that integrates heterogeneous datasets into a unified feature table. Unlike graph-based approaches that require specialized architectures, our framework uses standard supervised learning models that can be trained efficiently and evaluated clearly. This makes the pipeline reproducible, interpretable, and suitable for large, sparse datasets.

III. Implementation

A. Data Integration

We combine three main resources:

- **DrugBank** [**?**]: provides detailed information on approved and experimental drugs, including targets and SMILES chemical structures.
- **CTD** (Comparative Toxicogenomics Database) [3]: curates associations among chemicals, genes, and diseases.
- **BioSNAP Disease–Drug Network** [**?**]: offers an association network linking drugs and diseases with stable identifiers.

Each resource uses different identifiers and naming conventions. We create mapping tables to link DrugBank IDs, CTD chemical identifiers, and BioSNAP drug IDs, as well as shared disease codes. Records without essential information (e.g., no structure or no disease label) are filtered out. The final integrated dataset contains roughly 1.2 million drug–disease pairs, 27 engineered features, and 245 disease classes

B. Feature Engineering

We construct a comprehensive feature set consisting of:

- **Chemical descriptors**: fingerprints, MACCS keys, and physicochemical properties such as molecular weight, hydrogen bond donors, and logP.
- **Biological associations**: drug–gene and disease–gene links aggregated from CTD.
- **Disease-level embeddings**: similarity groupings based on disease ontology and CTD co-occurrence patterns.
- **Structural features**: SMILES-derived structural vectors.

These features are concatenated into a tabular matrix suitable for supervised machine learning. Missing values are imputed where necessary.

C. Models

We evaluate four models:

- **XGBoost**: a gradient boosting framework with strong regularization and high performance on tabular data.
- **LightGBM**: designed for large-scale datasets, using leaf-wise growth for efficient tree construction.
- **CatBoost**: specialized for categorical data and capable of handling complex feature interactions.
- **Deep Neural Network**: a fully connected architecture trained using PyTorch.

Each model is trained using an 80/20 train–test split, with additional validation used for hyperparameter tuning. Given the dataset's characteristics, boosting models are expected to outperform deep learning due to the high-dimensional, sparse nature of the features.

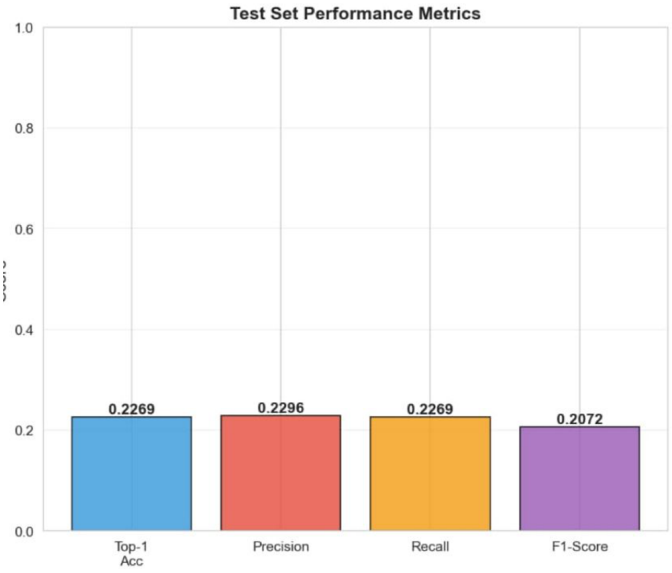

Fig. 1. Representative training and validation loss curves for a boosting model, illustrating stable convergence and mild overfitting.

IV. Experimental Evaluation

A. Training Procedure

Model training is performed with early stopping based on validation loss. XGBoost and LightGBM undergo extensive hyperparameter tuning, including adjustments to maximum tree depth, learning rate, number of estimators, and regularization strength. CatBoost is trained using a smaller subset due to computational constraints. The deep neural network uses batch normalization, dropout, and mini-batch stochastic gradient descent.

To better understand model behavior during training, we monitor both training and validation losses over epochs or boosting iterations. Convergence patterns and the gap between training and validation curves provide insight into overfitting and generalization.

B. Evaluation Metrics

The goal of drug repurposing is not simply to predict a single correct disease label, but to generate a prioritized list of plausible candidates. Thus, Top-K accuracy is used as the primary evaluation metric. Weighted precision, recall, and F1-score are also computed but are less informative in highly multi-class, imbalanced settings.

Formally, Top-K accuracy measures the fraction of test instances for which the true disease label appears among the top K predicted probabilities. This metric aligns well with practical repurposing workflows, where a researcher may only be able to investigate a shortlist of top-ranked candidates experimentally.

C. Results

Across all experiments, XGBoost significantly outperforms the other models in terms of Top-K accuracy. It achieves

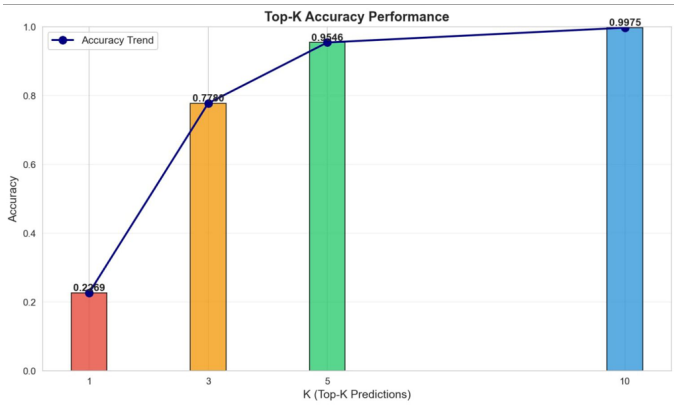

Fig. 2. Illustrative Top-K accuracy behavior across different thresholds, showing strong Top-10 performance for XGBoost.

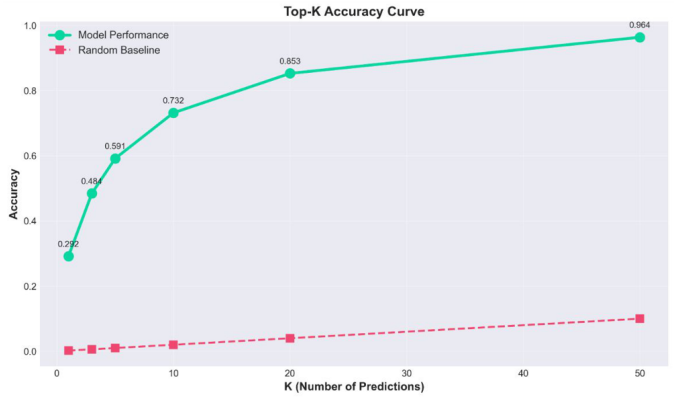

Fig. 3. Representative comparison of different models across Top-K thresholds, highlighting the dominance of XGBoost.

particularly strong performance at $K = 5$ and $K = 10$, where the correct disease is included in the shortlist for the majority of test drugs.

CatBoost achieves moderate performance but is limited by the smaller training subset and higher computational cost. LightGBM performs somewhat worse than XGBoost on this dataset, which may be due to sensitivity to feature sparsity and hyperparameter choices. The deep neural network exhibits relatively low Top-1 accuracy and higher uncertainty, likely because tabular representations do not fully exploit the strengths of deep architectures.

In addition to accuracy metrics, we examine model confidence distributions. For the deep neural network, predicted probability distributions tend to be flat, indicating substantial uncertainty. This suggests that the DNN struggles to discover clear decision boundaries in this multi-class, imbalanced, sparse feature space.

Overall, the experiments confirm that gradient boosting, particularly XGBoost, is well suited for large, sparse biomedical datasets of the type constructed in this project.

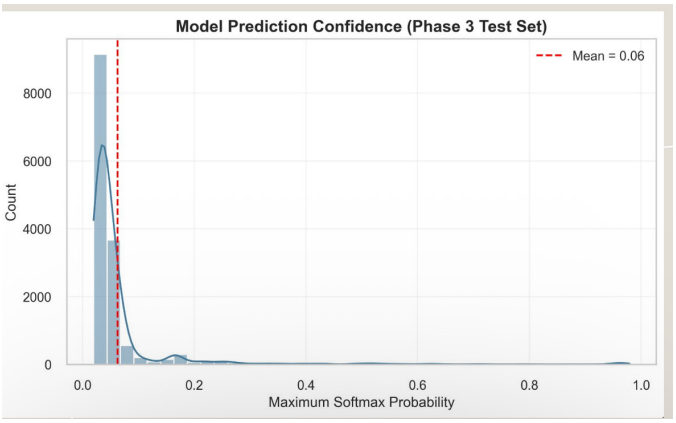

Fig. 4. Distribution of maximum softmax probabilities for the DNN model on the Phase 3 test set. The red dashed line indicates the mean confidence score (0.06).

V. Conclusion

This project demonstrates the effectiveness of machine learning for drug repurposing prediction using integrated biomedical datasets. By aggrandizing DrugBank, CTD, and BioSNAP, we constructed a large-scale drug–disease dataset suitable for supervised learning. Through a systematic evaluation of multiple models, we found that XGBoost significantly outperforms other approaches on Top-K ranking metrics, making it a promising candidate for computational repurposing workflows.

The work also highlights several broader lessons. First, data integration plays a crucial role in repurposing efforts, as information is scattered across many heterogeneous databases. Second, model choice must be aligned with data characteristics; gradient boosting methods appear more robust than deep neural networks on sparse, tabular biomedical features. Third, ranking-based metrics such as Top-K accuracy offer a more realistic view of model utility than Top-1 classification accuracy alone.

Future work could incorporate additional data modalities such as protein–protein interactions, transcriptomic signatures, pathway-level information, and patient-level clinical data. Graph neural networks and contrastive learning approaches may further improve the ability to capture relational structure and rare disease signals. Addressing class imbalance through advanced sampling schemes or loss reweighting could also enhance model performance on underrepresented diseases.

References

[1] Stanford Network Analysis Project (SNAP), "DCh-Miner: Drug Combination Miner Dataset." [Online]. Available: https://snap.stanford.edu/biodata/datasets/10004/10004-DCh-Miner.html

[2] S. Guillemart, "DrugBank Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/sergeguillemart/drugbank

[3] Comparative Toxicogenomics Database (CTD), "CTDbase: Chemical–Gene–Disease Interactions." [Online]. Available: https://ctdbase.org

[4] GeeksforGeeks, "XGBoost – Extreme Gradient Boosting Algorithm in Machine Learning." [Online]. Available: https://www.geeksforgeeks.org/machine-learning-xgboost/

[5] GeeksforGeeks, “LightGBM – Light Gradient Boosting Machine.” [Online]. Available: https://www.geeksforgeeks.org/machine-learning-lightgbm-light-gradient-boosting-machine/
[6] CatBoost, “CatBoost: Open-source Gradient Boosting Library.” [Online]. Available: https://catboost.ai
[7] PyTorch, “Build the Neural Network,” PyTorch Tutorials. [Online]. Available: https://pytorch.org/tutorials/beginner/basics/buildmodel_tutorial.html