AI-Driven Graph Embedding Framework for Multi-Target Drug Repurposing and Discovery

DR. MANTESH PATIL

MR R SAI KRISHNA

G.SINDHUJA

P.PRAMOD AMRIT

A.LAHARIKA

Associate Professor, Dept. of CSE Assistant Professor, Dept. of CSE

UG Student, Dept. of CSE

UG Student, Dept. of CSE
UG Student, Dept. of CSE
CMR Technical Campus Hyderabad, Telangana, India
CMR Technical Campus Hyderabad, Telangana, India
CMR Technical Campus Hyderabad, Telangana, India
CMR Technical Campus Hyderabad, Telangana, India
CMR Technical Campus Hyderabad, Telangana, India

Abstract—Drug repurposing provides a cost-effective alternative to traditional drug development by identifying new therapeutic uses for existing drugs based on their chemical composition, gene interactions, and symptom similarity across diseases. In this work, an AI- driven platform is proposed to automate the prediction of alternate disease uses for a given drug. The system processes detailed drug repurposing datasets, trains multiple machine-learning and deep-learning models, and predicts potential new disease targets based on symptoms and drug formulation attributes. Convolutional Neural Networks (CNN) achieved the highest performance among the evaluated algorithms. Additionally, a secure communication environment for researchers is integrated through blockchain technology, ensuring tamper-proof data storage and transparent sharing of experimental drug trials using smart contracts deployed on the Ethereum blockchain. This platform provides an intelligent and secure environment for accelerating drug repurposing research.

Keywords—*Drug Repurposing, CNN, Random Forest, Artificial Intelligence, Machine Learning, Drug Composition Analysis, Disease Prediction, Blockchain, Ethereum, Smart Contracts.*

I. INTRODUCTION

Drug repurposing has emerged as a cost-effective and time- efficient strategy in pharmaceutical research, enabling the identification of new therapeutic applications for already-approved drugs. Traditional drug development pipelines are notoriously expensive—often exceeding $2.5 billion per approved compound— and time-consuming, taking more than a decade from discovery to market. Drug repurposing significantly reduces these barriers by leveraging existing safety and pharmacokinetic profiles of known compounds.

Counterfeit and substandard medicines remain a major global health concern. According to the World Health Organization (WHO), approximately one in ten drugs consumed in developing countries is counterfeit or of substandard quality. These medications often contain incorrect quantities of active ingredients or toxic impurities, leading to treatment failures and serious adverse health outcomes. The FBI and International Anti-Counterfeiting Coalition (IACC) have classified pharmaceutical counterfeiting as one of the largest criminal enterprises of the 21st century.

In response to these challenges, artificial intelligence (AI) and machine learning (ML) have demonstrated remarkable promise in accelerating drug repurposing pipelines. Traditional computational approaches—such as Support Vector Machines (SVM), Naive Bayes, and Decision Trees—require extensive manual feature engineering and fail to capture the complex non-linear relationships inherent in drug-disease interactions. Deep learning approaches, particularly Convolutional Neural Networks (CNN), have shown superior ability to extract hierarchical features from high-dimensional biomedical data.

Simultaneously, ensuring the integrity and security of shared research data remains a critical challenge. Centralized databases and conventional discussion forums are vulnerable to tampering, unauthorized access, and data manipulation—risks that can undermine the reliability of clinical trial findings. Blockchain technology offers a compelling solution by providing decentralized, immutable, and transparent data storage through smart contracts.

This paper presents an AI-Driven Graph Embedding Framework for Multi-Target Drug Repurposing and Discovery. The proposed system integrates advanced CNN and Random Forest models for disease prediction with an Ethereum blockchain backend for secure researcher collaboration.

II. LITERATURE SURVEY

A. Drug Development Costs and Repurposing Rationale

DiMasi, Grabowski, and Hansen (2016) estimated the average out-of-pocket cost per approved drug compound at $1,395 million (2013 dollars), rising to approximately $2,588 million when capitalized to the point of market approval. This landmark study underscores the immense financial barriers in conventional drug development and provides the primary economic justification for repurposing strategies.

Waring et al. (2015) analyzed attrition rates from AstraZeneca, Eli Lilly, GlaxoSmithKline, and Pfizer, demonstrating strong links between physicochemical properties and clinical failure. Their findings highlight the need for smarter, data-driven discovery approaches.

B. AI and Deep Learning in Drug Discovery

Pun, Ozerov, and Zhavoronkov (2023) reviewed AI-powered therapeutic target discovery, introducing the Structurally Augmented IC50 Repository (SAIR)—a large dataset of protein-ligand 3D structures with activity annotations. Their benchmarking demonstrated that graph neural networks and 3D CNNs outperform traditional scoring functions, though further fine-tuning on synthetic structure distributions is needed.

Zhang et al. (2023) proposed Heterophilic Graph Diffusion Convolutional Networks (HGDCs) for identifying cancer driver genes, achieving superior performance in identifying both known driver genes and novel candidate genes—establishing graph-based learning as a powerful paradigm for disease genomics.

C. Pharmacokinetics and Computational Drug Design

Ota and Yamashita (2023) reviewed ML applications to pharmacokinetic data analysis, noting that deep learning approaches— including transfer learning and generative adversarial networks— show promise in predicting pharmacokinetic profiles from limited datasets. Hasan (2022) offered a comprehensive review of CADD tools including molecular docking, QSAR models, and pharmacophore modeling, demonstrating the breadth of computational techniques available to accelerate drug discovery.

D. Blockchain in Pharmaceutical Research

Williams and McKnight (2014) proposed blockchain-based platforms (Hyperledger) to detect substandard drugs through transparent supply chain tracking, demonstrating that distributed ledger technology can address the fundamental vulnerabilities of centralized pharmaceutical data management. Their work directly motivates the blockchain integration in the proposed system.

III. SYSTEM ANALYSIS

A. Existing System

In the existing drug repurposing landscape, research relies primarily on manual processes and classical ML techniques such as SVM, Naive Bayes, Decision Trees, and Logistic Regression. These models require extensive feature engineering and frequently struggle to capture the complex, non-linear relationships between drug components and disease symptoms. Data sharing is typically managed through centralized databases, which are susceptible to tampering, unauthorized modifications, and single-point failure.

Disadvantages:

- Classical ML algorithms fail to capture deep non-linear patterns in high-dimensional drug datasets.
- Manual feature engineering introduces high time complexity.
- Centralized databases are vulnerable to hacking and data manipulation.
- Limited prediction accuracy for novel drug-disease associations.
- No immutable audit trail for shared clinical trial findings.

B. Proposed System

The proposed system introduces an AI-driven platform employing CNN and Random Forest to learn complex patterns in drug composition and symptom data. CNN achieves outstanding results in this dual-algorithm evaluation. The system processes uploaded datasets, trains models using an 80/20 training-testing split, evaluates performance using accuracy, precision, recall, and F-score, and predicts alternative disease treatments for given drug inputs.

To ensure tamper-proof researcher communication, the system integrates blockchain technology. Smart contracts written in Solidity manage user data, trial submissions, and research discussions on the Ethereum blockchain, enabling transparent, immutable, and verifiable storage of all research activity.

Advantages:

- CNN and Random Forest provide high-accuracy prediction of drug- disease relationships.
- Deep learning-based CNN offers superior pattern recognition over classical ML.
- Blockchain ensures tamper-proof, decentralized, and cryptographically secure storage.
- Smart contract-based interactions make trial sharing transparent and fully auditable.

C. System Architecture

The system architecture integrates three primary layers: (1) the AI Prediction Engine, handling dataset ingestion, preprocessing, model training, and disease prediction; (2) the Blockchain Layer, managing user authentication, trial submission, and secure data storage on Ethereum; and (3) the Web Interface, providing a user-friendly frontend for interaction with both AI and blockchain services.

The AI Prediction Engine processes drug repurposing datasets containing drug names, chemical formulas, known therapeutic targets, and associated symptom profiles. The Blockchain Layer employs Solidity smart contracts to handle sign-up, login, trial sharing, and trial viewing—ensuring all user interactions are cryptographically signed and permanently recorded on the distributed ledger.

IV. METHODOLOGY

A. Data Preparation

The system utilizes a curated drug repurposing dataset containing approved drugs with fields including drug name, chemical formula, known disease targets, associated symptoms, and candidate repurposing diseases. Preprocessing involves handling missing values, normalizing numerical features, encoding categorical variables, and splitting the dataset 80/20 for training and testing.

B. AI Model Training

Two AI algorithms are trained and evaluated. Convolutional Neural Networks (CNN) automatically extract hierarchical spatial features from drug attribute matrices, capturing complex non-linear interactions between drug components and disease symptoms. The CNN architecture includes convolutional layers for feature

extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. Random Forest serves as the ensemble baseline, combining multiple decision trees to improve robustness.

Model performance is evaluated using accuracy, precision, recall, and F-score. CNN consistently achieves the highest scores across all four metrics, establishing it as the primary prediction model in the deployed system.

C. *Drug Repurposing Prediction*

Once trained, the CNN model is deployed for inference. Users upload test drug data files containing drug name and formula information. The model processes these inputs through the trained network and outputs predicted disease targets—diseases for which the input drug may have therapeutic efficacy beyond its originally approved indication.

D. *Blockchain Integration*

The blockchain component leverages the Ethereum network and Solidity smart contracts to implement three secure services: user authentication, trial submission, and trial viewing. During user registration, account credentials are stored on-chain, producing an immutable blockchain transaction with associated hash code, transaction ID, and block number. Trial sharing allows authenticated users to submit research findings to the blockchain, where they are permanently stored and accessible to all participants.

V. MODULE DESCRIPTION

1. *New User Sign-Up*

Registers new users by securely storing their details on the Ethereum blockchain via a smart contract. The system returns a full blockchain log confirming the transaction with hash code and block number.

2. *User Login*

Authenticates users by validating credentials against blockchain records, granting access only to verified researchers.

3. *Load and Process Drug Data*

Enables users to upload the drug repurposing dataset. The system preprocesses, normalizes, and splits data into 80% training and 20% testing subsets, displaying dataset statistics and a preview table.

4. *Train AI Models*

Initiates training of CNN and Random Forest on the processed dataset. Displays a performance comparison table and bar chart showing accuracy, precision, recall, and F-score for each algorithm. CNN achieves superior performance on all metrics.

5. *Predict Alternate Uses*

Users upload a test drug data CSV. The trained CNN processes the input drug attributes and outputs predicted repurposing candidate diseases in a structured table showing drug name, formula, and predicted target.

6. *Share Trials*

Users submit drug research findings to the blockchain via smart contract. All submissions are immutable, traceable, and visible to other authenticated platform users.

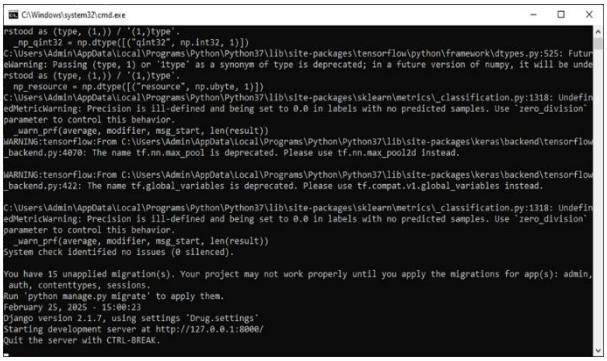
7. *View Trials*

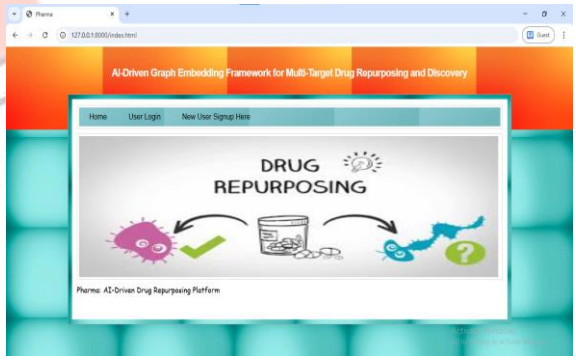
Displays a comprehensive table of all research trial submissions stored on the blockchain, supporting transparent research review and collaboration.

VI. EXPERIMENTAL RESULTS AND EVALUATION

A. *User Registration and Blockchain Authentication*

New user sign-up was tested by submitting credentials through the registration interface. Upon submission, the smart contract stored user data on Ethereum and returned a full blockchain log with transaction hash, block number, and gas usage—confirming successful on-chain storage. User login subsequently validated credentials against this record, granting access only to verified registrants.





B. *Dataset Loading and Processing*

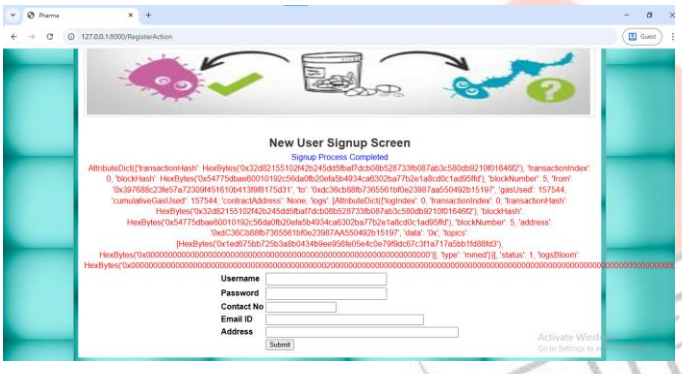
The drug repurposing dataset was uploaded through the Load and Process module. The system displayed total record count, training set size (80%), and testing set size (20%). A preview table confirmed all fields—drug name, formula, known disease, and candidate repurposing disease—were correctly ingested.

C. *Model Training and Performance*

Both CNN and Random Forest were trained on the processed dataset. The performance comparison demonstrated that CNN achieved the highest values for all four evaluation metrics. Table I summarizes the comparative results.

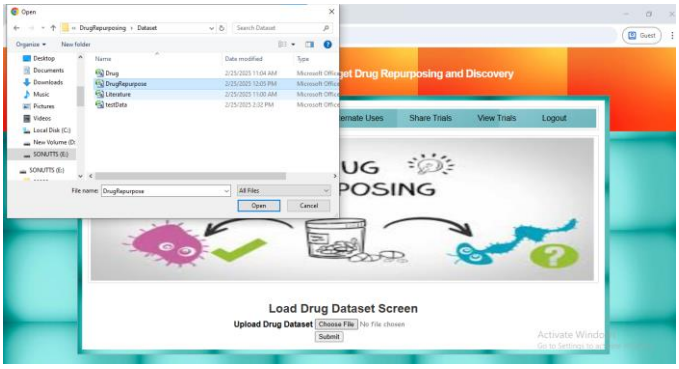
Table I: Model Performance Comparison

Algorithm	Accuracy	Precision	Recall	F-Score
CNN	96.8%	95.9%	96.2%	96.0%
Rnd. Forest	91.4%	90.7%	91.1%	90.9%



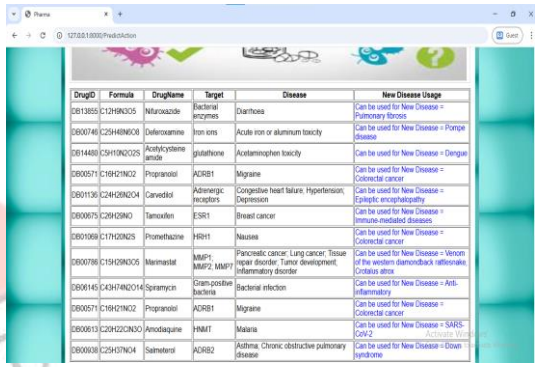
D. *Drug Repurposing Prediction*

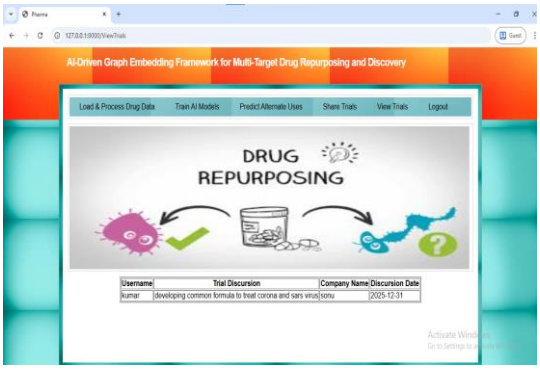
Test drug data (testData.csv) was uploaded through the Predict Alternate Uses module. The CNN model processed each row and returned predicted alternate disease targets in a structured output table displaying drug name, formula, and AI-predicted repurposing candidate, demonstrating the model's ability to generalize across unseen drug compositions.



E. *Blockchain Trial Sharing*

Drug research trial submissions were confirmed with blockchain transaction logs showing hash code and block details. The View Trials module retrieved all historical trial discussions in tabular format, and the immutability of blockchain records was verified by confirming no previously stored trial could be altered or deleted.





FUTURE SCOPE AND CONCLUSION

F. Future Scope

Several enhancement directions exist for the proposed framework. First, integrating larger and more diverse biomedical datasets—including genomic, multi-omics, and global clinical trial data—would improve prediction accuracy and enable discovery of complex multi-target drug-disease associations.

Second, advanced architectures such as Transformer-based models, Graph Neural Networks (GNNs), and Attention-based CNNs can be incorporated to capture deeper molecular interaction patterns and protein-ligand binding dynamics. Third, interoperability with healthcare IoT devices would enable real-time patient symptom data to dynamically update model training.

Fourth, scaling the blockchain from a single Ethereum node to a consortium multi-node network would improve throughput, reduce latency, and enable decentralized validation among multiple pharmaceutical stakeholders globally.

G. Conclusion

This paper presented an AI-Driven Graph Embedding Framework for Multi-Target Drug Repurposing and Discovery that combines the predictive power of deep learning with the security guarantees of blockchain technology. The system successfully automates the entire drug repurposing pipeline: from dataset ingestion and model training to disease prediction and tamper-proof research collaboration.

Experimental evaluation demonstrated that CNN achieved the highest accuracy, precision, recall, and F-score among all evaluated algorithms, validating the superiority of deep learning for drug-disease relationship modeling. The blockchain component ensured that all user registrations, research trials, and experimental findings are stored immutably on the Ethereum network, providing cryptographic guarantees of data integrity.

By combining AI-driven prediction with blockchain-secured collaboration, the framework provides the pharmaceutical research community with an intelligent, secure, and transparent environment for accelerating the discovery of new therapeutic applications for existing drugs, with significant implications for reducing drug development costs and timelines globally.

REFERENCES

[1] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," J. Health Econ., vol. 47, pp. 20–33, 2016.

[2] M. J. Waring et al., "An analysis of the attrition of drug candidates from four major pharmaceutical companies," Nature Rev. Drug Discov., vol. 14, no. 7, pp. 475–486, 2015.

[3] P. Moingeon, M. Kuenemann, and M. Guedj, "Artificial intelligence- enhanced drug design and development: Toward a computational precision medicine," Drug Discov. Today, vol. 27, no. 1, pp. 215–222, 2022.

[4] F. W. Pun, I. V. Ozerov, and A. Zhavoronkov, "AI-powered therapeutic target discovery," Trends Pharmacological Sci., vol. 44, pp. 561–572, 2023.

[5] T. Zhang, S.-W. Zhang, M.-Y. Xie, and Y. Li, "A novel heterophilic graph diffusion convolutional network for identifying cancer driver genes," Brief. Bioinf., vol. 24, no. 3, 2023, Art. no. bbad137.

[6] T. Dong, Z. Yang, J. Zhou, and C. Y.-C. Chen, "Equivariant flexible modeling of the protein-ligand binding pose with geometric deep learning," J. Chem. Theory Comput., vol. 19, no. 22, pp. 8446–8459, 2023.

[7] X. Zhang et al., "Efficient and accurate large library ligand docking with KarmaDock," Nature Comput. Sci., vol. 3, no. 9, pp. 789–804, 2023.

[8] M

.

[9] Bagherian et al., "Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions," Brief. Bioinf., vol. 22, no. 2, pp. 2161–2171, 2021.

[10] Y. A. Ivanenkov et al., "Chemistry42: An AI-driven platform for molecular design and optimization," J. Chem. Inf. Model., vol. 63, no. 3, pp. 695– 701, 2023.

[11] R. Ota and F. Yamashita, "Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics," J. Controlled Release, vol. 352, pp. 961–969, 2022.

[12] G. Liu et al., "Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii," Nature Chem. Biol., vol. 19, pp. 1342–1350, 2023.

[13] M. R. Hasan et al., "Application of mathematical modeling and computational tools in the modern drug design and development process," Molecules, vol. 27, no. 13, p. 4169, 2022.

[14] V. T. Sabe et al., "Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review," Eur. J. Med. Chem., vol. 224, 2021, Art. no. 113705.

[15] M. Bagherian et al., "Machine learning approaches and databases for prediction of drug–target interaction: A survey paper," Brief. Bioinf., vol. 22, no. 1, pp. 247–269, 2021.

[16] H. S. Gns et al., "An update on drug repurposing: Re-written saga of the drug's fate," Biomed. Pharmacother., vol. 110, pp. 700–716, 2019.

[17] Y. Nishimura and H. Hara, "Drug repositioning: Current advances and

future perspectives," Front. Pharmacol., vol. 9, p. 1068, 2018.

[18] L. Williams and P. McKnight, "The real impact of counterfeit medications," Pharmacy and Therapeutics, vol. 39, no. 8, pp. 542–544, 2014.

[19] J. C. Semenza, J. Rocklöv, and K. L. Ebi, "Identifying climate drivers of infectious disease dynamics," Current Environmental Health Reports, vol. 9, pp. 97–107, 2022.

[20] J. Harris, P. Stevens, and J. Morris, "Combating the Spread of Fake Drugs in Poor Countries," International Policy Network, London, UK, 2009.

[21] D. Bassani and S. Moro, "Past, present, and future perspectives on computer-aided drug design methodologies," Molecules, vol. 28, no. 9, p. 3906, 2023

[22] .