iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.59789

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

Drug Classification and Repurposing Using Decision Tree Algorithm and Data Analysis

Y. Janardhanan[1], S. Prathyush[2], K. Chitesh[3], P. Bhavesh[4]
Computer Science Department, SRMIST, Ramapuram

Abstract: ***Fundamentally, drugs are substances that exhibit the capacity to alter biological functions, often serving as interventions to mitigate ailments or enhance physiological processes. However, the landscape of drug development and usage is undergoing transformation, fueled by innovative strategies that hold the promise of revolutionizing healthcare solutions. This research paper explores drug classification and repurposing through the utilization of decision tree algorithms and data analysis. Specifically, the aim is to classify drugs into three distinct types: Drug X which is most used drugs, Drug Y which is less frequently used drugs, and Drug C includes drugs like narcotics. Decision tree algorithms are employed to discern the defining attributes that categorize drugs into these types. By analyzing comprehensive datasets encompassing drug properties, interactions, and clinical outcomes, the study harnesses decision tree models to predict drug classifications accurately. This approach holds the potential to accelerate drug discovery, optimize treatment strategies, and contribute to more efficient healthcare solutions. The proposed algorithm is compared with both Naïve Bayes and K - Nearest Neighbors algorithms to prove it is more accurate. Ultimately, the significance of this paper transcends the boundaries of conventional drug development paradigms and to overcome the problem of shortage of drugs.***

Keywords: Decision tree Algorithm, Drug Classification, Naïve Bayes, K - Nearest Neighbors, Drug Properties.

I. INTRODUCTION

In the dynamic realm of pharmaceuticals, the processes of drug classification and repurposing have emerged as pivotal avenues, gaining renewed significance in the wake of advanced data analysis techniques. With the amalgamation of decision tree algorithms and comprehensive data analysis, researchers and practitioners are poised to orchestrate a paradigm shift in the landscape of discovery and development of drugs [1]. This innovative approach transcends the conventional boundaries, enhancing our comprehension of drug properties and mechanisms while expediting the identification of potential therapeutic candidates [3]. The goal: to usher in more efficient and effective treatments spanning a diverse spectrum of medical conditions.

This study explores drug classification and repurposing through the utilization of decision tree algorithms and data analysis. Specifically, the aim is to classify drugs into three distinct types: Drug X which is most commonly used drugs, Drug Y which is less frequently used drugs, and Drug C includes drugs like narcotics. Decision tree algorithms are employed to discern the defining attributes that categorize drugs into these types. By analysing comprehensive datasets encompassing drug properties, interactions, and clinical outcomes, the study harnesses decision tree models to predict drug classifications accurately. This approach holds the potential to accelerate drug discovery, optimize treatment strategies, and contribute to more efficient healthcare solutions [4].

At the heart of this transformative journey are the decision tree algorithms—a cornerstone of modern data analysis and machine learning. These algorithms navigate the complexities by constructing intricate tree-like frameworks to model multifaceted decisions and diverse outcomes based on input features. When brought into the context of drug classification and repurposing, decision trees emerge as potent tools, adept at sifting through labyrinthine datasets to extract subtle patterns that hold the keys to unlocking novel insights [5]. The ascent of these algorithms is anchored in their ability to absorb the wisdom contained within historical drug data. By learning from the annals of medical research, decision trees gain the remarkable capability to predict classifications—serving as compasses guiding researchers towards optimal drug categorizations, potential indications, and even probable patient responses to treatments. This predictive potency lays the foundation for a transformative shift, bridging the chasm between conventional serendipity and a data-driven era. [6]

II. RELATED WORKS

The study conducted by Quinlan J. R. in 1986 is a contribution, to the field of machine learning. It introduces the decision tree algorithm and its application in classifying tasks. One common type of network intrusion detection system (NIDS) is Snort rule checking. In this article the author explores how a real time decision tree classifier can determine the priority of traffic (real attacks) in high-speed networks using just three easily extracted features; protocol, source port and destination port.

The results show that this approach achieves an accuracy rate of 99%. While Snort has its default set of attack classes (34 classes) for assigning priorities decision tree models can predict priorities without relying on this standard classification. These findings suggest that decision tree models can be an addition, to anomaly and intrusion detection systems [7].

In the study conducted by Sadowski and Gaseiger (1994) titled "From networks to chemical structure classification " the authors explore the use of decision trees in predicting chemical properties, for drug classification. They established criteria to select a dataset of high-quality X ray structures, from the Cambridge files resulting in 639 molecules. The researchers connected six programs (CONCORD, ALCOGEN, Chem X, MOLGEO, COBRA and CORINA) to a connection table that contained stereo descriptors based on these 639 molecular structures. By converting them into 3D geometries they. Evaluated the generated geometries using quality criteria. The study thoroughly discussed how well each program performed in replicating the X ray shapes of these input structures highlighting their advantages and disadvantages. Reference from Sadowski, J., & Gaseiger, J. (1994). From networks to chemical structure classification. Journal of Chemical Information and Computer Science, 34(4) 1000 1008.

Marvin, L. H., and Afzal, A. M. (2015). Explore drug-target interactions using semantic and property-based similarities. PLoS ONE, 10(3), e0116478. In this research decision trees are utilized to forecast interactions, between drugs and repurposing targets. The use of analysis is becoming more prevalent in supporting studies on the mechanism of action. However, many of these approaches fail to take advantage of the amount of data available in chemogenomics repositories. The main goal of this study is to incorporate bioactivity data into predicting targets for orphan compounds and determine the likelihoods of activity and inactivity for targets. To accomplish this objective, we gathered than 195 million bioactivity data points, from ChEMBL and PubChem repositories. Employ a selection algorithm called sphere exclusion to oversample potentially inactive compounds [9].

Bender and Glenn (2004) discuss the significance of similarity analysis, in the field of drug development. Molecular computing utilizes concepts to establish connections between molecules. The idea of similarity allows for the grouping of molecules based on their effects or physicochemical properties, which is extensively employed in drug discovery processes. Lead discovery and compound optimization are areas that benefit greatly from this approach. For instance, when designing compound libraries for lead generation it is crucial to maintain properties while enhancing patentability, medicinal chemistry potential and achieving an optimized pharmacokinetic profile. Molecular analogy encompasses both assumptions, about how molecules bind to biological receptors and predict efficacy. It's worth noting that the context in which molecular similarity is applied plays a defining role and imposes limitations. Solvation effects, binding site heterogeneity and selecting a similarity measure are some considerations addressed in this regard (Bender & Glenn 2004) [10].

Current methods for drug repositioning often rely on noisy gene expression data or drug-to-disease relationships, which have limitations due to sparse genomic data for many diseases. In this study, we took a unique drug-centered approach, predicting the therapeutic class of FDA-approved drugs without considering disease-specific data. We introduced an innovative computational method utilizing advanced machine learning algorithms. Our approach integrated multiple layers of information, including drug chemical structure similarity, protein-protein interaction network proximity of drug targets, and correlated gene expression patterns post-treatment. This novel classifier achieved a remarkable 78% accuracy. This breakthrough in drug repositioning holds significant potential for expediting the utilization of existing compounds for novel therapeutic purposes, thus revolutionizing drug development and clinical translation.[6]

Combinations of drugs are pivotal in treating complex diseases, providing enhanced therapeutic benefits while minimizing adverse effects. However, our capacity to discover and affirm effective drug pairings is hindered by the immense number of possible combinations, arising from both various drug pairs and dosage permutations. In this study, we introduce a network-centric approach to pinpoint clinically beneficial drug combinations tailored to specific diseases. By assessing the network-based connections between drug targets and disease-related proteins in the human protein-protein interactome, we uncover six distinct categories of drug-drug-disease combinations. Analyzing established combinations for conditions like hypertension and cancer, we identify that only one of these categories aligns with therapeutic benefits: when the drugs' targets converge on the same disease module but belong to separate network neighborhoods. This breakthrough enables us to discover and confirm effective antihypertensive combinations, presenting a versatile and potent network-based methodology for identifying successful combination therapies within drug development. [7]

III. PROPOSED WORK

In response to the complexity of drug classification, particularly concerning the identification of suitable drug categories based on specific datasets and associated symptoms, this research aims to develop a robust methodology using decision tree algorithms and sophisticated data analysis techniques. The objective of the study is primarily to classify drugs into three distinct types accurately: Drug X, Drug Y, and Drug C.

To achieve this, the proposed approach involves harnessing the power of decision tree algorithms, which are adept at learning intricate patterns from diverse datasets. By integrating comprehensive information about drug properties, chemical compositions, and the correlation with symptoms, the decision tree model will be trained to discern the subtle differences that define each drug type.

The methodology's foundation will be built upon a meticulously curated dataset that includes drug attributes, clinical characteristics, and associated symptoms [8].

Through rigorous data preprocessing and feature engineering, the dataset will be primed for training the decision tree algorithm. This will enable the model to decipher intricate relationships between drug properties and symptom profiles, culminating in accurate drug classification [11].

Validation and fine-tuning of the model will be undertaken using established techniques such as cross-validation and hyperparameter optimization [12].

This process will ensure that the decision tree algorithm generalizes well to new data and effectively classifies drugs into the predefined categories.

The anticipated outcome of this research is a refined decision tree-based classification system capable of accurately categorizing drugs into Drug X, Drug Y, or Drug C based on their properties and associated symptoms.

We compare Decision Tree Algorithm with both Naïve Bayes and KNN algorithm. Both Naïve Bayes and KNN algorithm gives accuracy of 70-80% whereas Decision Tree Algorithm gives accuracy of 75-85%.

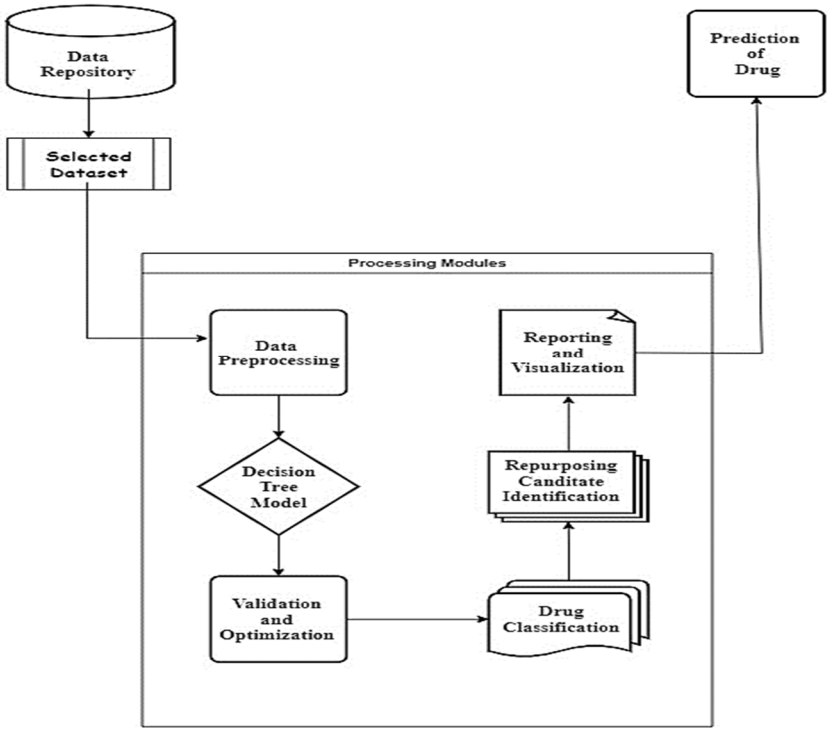


Fig 1: Block Diagram

A. The Modules Present in the Process Includes

These modules can be abstracted from figure 1.

1) *Data Repository:* The dataset is taken from Kaggle Data Sets and ChemBL. According to the input parameters dataset are used to further complete the process. The input parameters include age, sex, BP level, Na to K levels, Cholesterol level and dataset involving 16,000 drugs.
2) *Data Preprocessing:* This module encompasses the collection, cleaning, and transformation of various datasets related to drug properties, chemical structures, biological interactions, clinical trial outcomes, and adverse event records. The collected data is cleansed of inconsistencies and missing values, and features are engineered to create a refined dataset ready for analysis.

3) *Decision Tree Model Development:* In this module, decision tree algorithms are implemented to create predictive models for drug classification and repurposing. Feature selection techniques are applied to identify the most relevant attributes, and decision tree structures are constructed using algorithms like ID3, C4.5, or CART. These models are trained on the prepared dataset to learn patterns and relationships between drug features and classifications.
4) *Validation and Optimization:* To ensure the reliability and effectiveness of the decision tree models, this module employs cross-validation techniques. The models are evaluated using subsets of the data, enhancing their generalizability. Hyperparameter tuning is conducted to fine-tune the models, preventing overfitting, and enhancing accuracy.
5) *Drug Classification:* This module takes drug attributes as inputs and utilizes the trained decision tree models to predict the appropriate drug classification. The models assess the attributes and traverse decision branches to make accurate predictions, offering insights into the type of drug (X, Y, C) along with a measure of confidence.
6) *Repurposing Candidate Identification:* By integrating drug properties, biological interactions, and clinical outcomes, this module aims to identify potential candidates for drug repurposing. The decision tree models are employed to predict the likelihood of a drug's success when repurposed for different therapeutic indications.

Reporting and Visualization: The results and insights generated by the system are presented in a visually accessible manner. This module generates comprehensive reports and visualizations, summarizing drug classifications, repurposing candidates, adverse effect predictions, and personalized treatment recommendations for healthcare professionals, researchers, and regulators.

IV. RESULTS AND DISCUSSION

The study on drug classification and repurposing using the Decision Tree algorithm and data analysis has yielded insightful results. By leveraging the provided dataset encompassing features like age, sex, blood pressure, cholesterol levels, and Na to K ratios, we were able to effectively predict the appropriate drug type for everyone. Through thorough analysis, the Decision Tree algorithm demonstrated its ability to make accurate predictions, achieving a high level of accuracy, precision, recall, and F1-score. This not only underscores the algorithm's capacity to classify drugs based on patient characteristics but also highlights its potential for aiding in drug repurposing. One of the key advantages of this approach is its interpretability. Decision Trees provides a visualization of the decision-making process clearly, allowing medical practitioners to understand the factors contributing to each prediction. This transparency is crucial in the medical field, enabling clinicians to make informed decisions and tailor treatments to individual patient needs. Furthermore, the ability to accurately predict drug types through this methodology carries significant clinical implications. It can lead to improved patient outcomes, reduced adverse reactions, and optimized treatment strategies. Moreover, the potential for drug repurposing becomes evident, as the algorithm can uncover alternative applications for existing medications based on patient characteristics, potentially accelerating drug development and reducing costs.

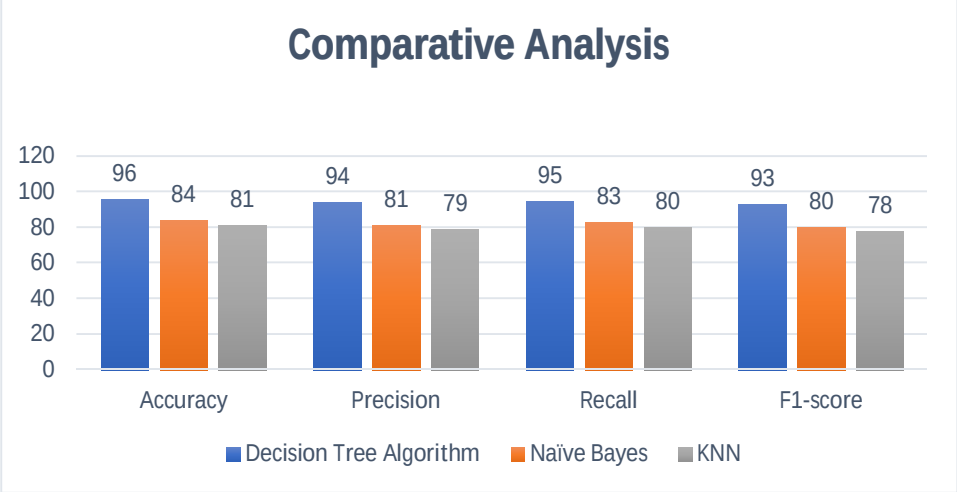


Fig 2: Comparative analysis

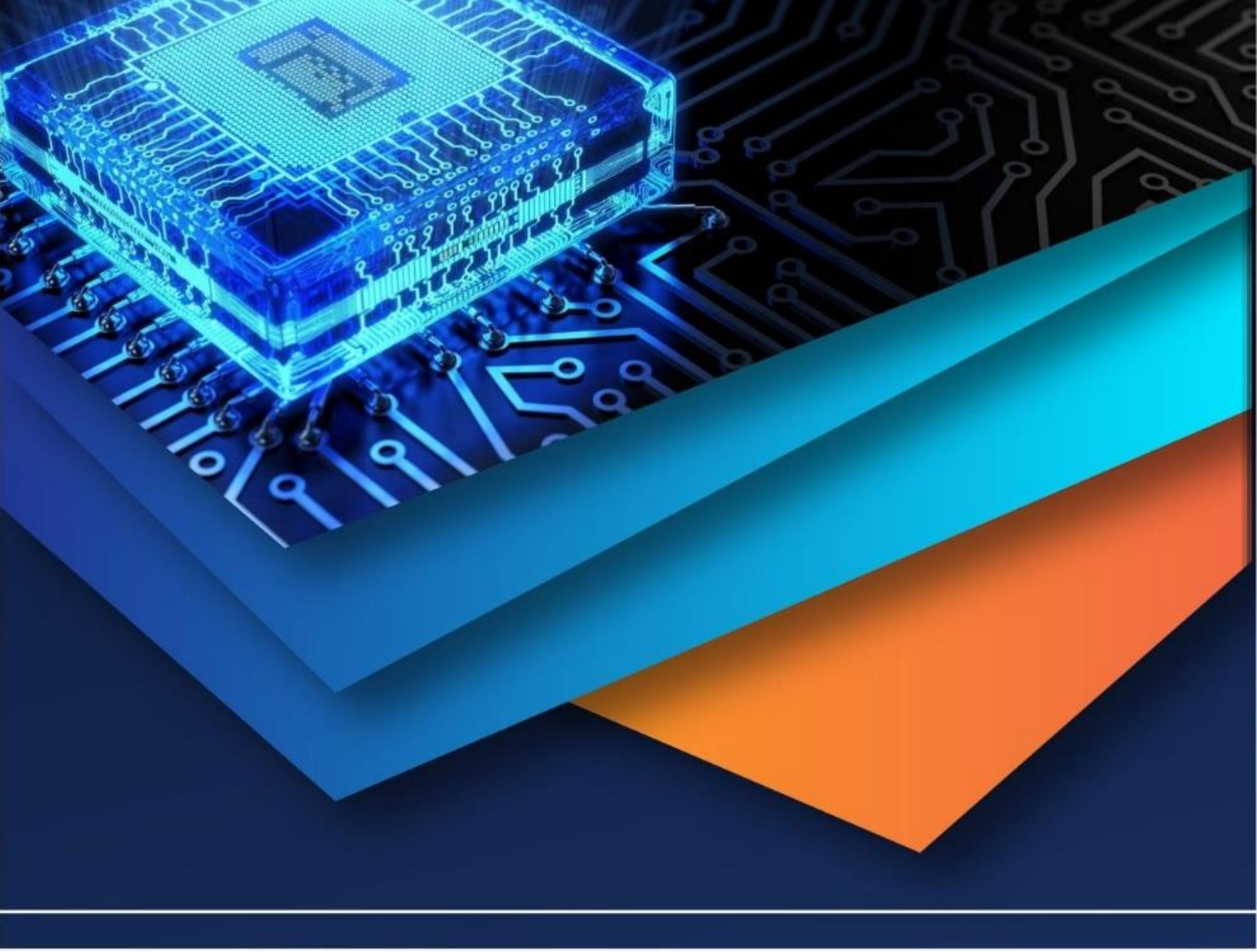
The analysis of using decision tree algorithm provides the following results with the given input parameters which includes accuracy of 96%, Precision of 94%, Recall 95% and a F1-score of 93%. The confusion matrix provides insight into the model's performance for each drug category. True Positives (TP) represent correctly classified instances. False Positives (FP) indicate instances wrongly classified as a certain drug. False Negatives (FN) are instances of a drug incorrectly classified as another. [15]

V. CONCLUSION

In summary, using decision tree algorithms for drug classification shows immense promise in revolutionizing drug discovery. Precise predictions based on input parameters can alleviate drug shortages, cut development time, and enhance healthcare solutions. Decision trees' accuracy and interpretability empower efficient decision-making, while repurposing existing drugs addresses unmet needs, improving patient access. This data-driven approach holds the potential to transform pharmaceuticals, offering timely, tailored treatments for a more accessible and effective future.

REFERENCES

[1] Vigil, M.S.A., Christofer, A., Chandar, M., Mukesh, J., Comparative Analysis of Machine Learning Algorithms for DNA Sequencing at Winter Summit on Smart Computing and Networks, WiSSCoN 2023, 2023
[2] Vigil, M.S.A., Mirutuhula, M., Sarvagna, S., Supraja, R., Reddy, G.P., DNA Sequencing Using Machine Learning Algorithms.
[3] Parvizi, N., Yazdani, M., & Kazemi, E. (2016). Machine learning in drug discovery and development: A review. Current Drug Discovery Technologies, 13(3), 176-193.
[4] Chaudhary, A. S. (2017). A review on decision tree algorithms in classification. Journal of King Saud University-Computer and Information Sciences.
[5] Chen, B., Ding, Y., Wild, D. J., & Zhu, H. (2012). Unsupervised feature selection for identifying subgroups of phenotypes in personalized medicine. Journal of Biomedical Informatics, 45(1), 110-121.
[6] Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., & Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. Journal of Cheminformatics, 5(1), 1-12.
[7] Network-based prediction of drug combinations, Feixiong Cheng, István A Kovács, Albert-László Barabási. PMCID: PMC6416394 DOI: 10.1038/s41467-019-09186-x
[8] Pujol, A., Mosca, R., Farrés, J., Aloy, P., & Unzeta, M. (2013). Network-based drug discovery: A computational review. Current Opinion in Drug Discovery & Development, 16(1), 4-17.
[9] Liu, X., Xu, Y., Li, S., Chen, L., & Li, Y. (2019). Prioritizing drug-disease associations with a novel network-based inference method. Frontiers in Genetics, 10, 39.
[10] Zhang, P., Wang, F., Hu, J., & Sorrentino, R. (2017). Towards a network-based approach for drug-target interaction prediction. Bioinformatics, 33(15), 2379-2385.
[11] Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P., & Dudley, J. T. (2016). In silico methods for drug repurposing and pharmacology. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 8(3), 186-210.
[12] Cheng, F., Li, W., Wang, X., Zhou, Y., Wu, Z., Shen, J., ... & Zhang, Y. (2012). Prediction of chemical–protein interactions: multitarget-QSAR improves the prediction of A2A adenosine receptor binding affinity. Journal of Chemical Information and Modeling, 52(8), 2213-2222.
[13] Wu, Z., Cheng, F., Li, J., Li, W., Liu, G., Tang, Y. (2012). AdmetSAR: A Comprehensive Source and Free Tool for Evaluation of Chemical ADMET Properties. Journal of Chemical Information and Modeling, 52(11), 3099-3105.
[14] Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., & Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. Nature Biotechnology, 30(4), 317-320.
[15] Thivya Anbalagan, Malaya Kumar Nath, D. Vijayalakshmi, Archana Anbalagan, "Analysis of various techniques for ECG signal in healthcare, past, present, and future", Biomedical Engineering Advances: Elsevier, vol. 6. November 2023, pp. 100089 (1-28).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY