Machine Learning Techniques for Pharmaceutical Bioinformatics

تقنيات التعلم الآلي في مجال المعلوماتية الحيوية الصيدلانية

by

AHMED ATTA AHMED SULTAN

A dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)

at

The British University in Dubai

November 2018

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

_____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

This dissertation presents a novel drug classifier to automate the prediction of drug indication and drug interactions with other drugs. The study integrates knowledge visualization, analysis, as well as development of a predictive model based on the Drug-Drug Interactions (DDIs) as a complex network. DDIs network analysis reveals unique drug features and explains unknown drug behaviors. Each drug molecule has a unique chemical structure and a set of pharmacological features. This set of attributes imposes how each drug performs its action inside a human body. Drug molecule interacts with multiple components in the biological system, for example, enzymes, proteins, among other drugs. The complexity of the chemical and pharmacological features forces the interaction between drug molecule and all other entities in the biological system to follow specific rules. The full features for each drug are not fully explained by researchers due to the incomplete drug profile description. DDIs network has a significant role in drug repurposing; it uncovers the hidden properties of the drug behavior. Predicting drug properties is presented as a contribution effort to drug repositioning approach.

To confirm the visual analysis, a binary matrix is drawn from each drug profile based on DDIs dataset. In this matrix, each drug is represented by a vector of attributes from all other drugs. A predictive model is developed to predict drug indication as well as to predict new DDIs using multiple machine learning algorithms.

This dissertation presents a case study of predicted anti-cancer activity for 38 drugs. The proposed Artificial Intelligence approach for drug-related properties prediction

demonstrates a high potential in complementing the current computational techniques. The predicted anti-cancer activity is computationally validated by a 10-fold cross validation evaluation technique and clinically supported by extensive literature review confirming the achieved results. In conclusion, the predicted drug features can provide new directions towards promising candidates for drug repositioning.

الملخص:

يعرض هذا البحث مصنفاً جديداً للعقاقير للتنبؤ بخصائص الأدوية وكذلك تفاعلات الأدوية مع بعضها البعض. تشمل الدراسة التصور المعرفي والتحليل ، بالإضافة إلى تطوير نموذج التنبؤ القائم على التفاعلات الدوائية-الدوائية كشبكة معقدة .تحليل هذه الشبكة يكشف عن خصائص الدواء الفريدة ويوضح بعض الخصائص الغير معروفة. كل جزيء دواء له تركيبة كيميائية فريدة ومجموعة من الخصائص الدوائية .هذه الصفات تحدد كيف يقوم كل دواء بعمله داخل جسم الإنسان. يتفاعل جزيء الدواء مع مكونات متعددة في النظام البيولوجي على سبيل المثال إنزيمات وبروتينات بالإضافة إلى عقاقير أخرى. إن تعقيد السمات الكيميائية والدوائية يجبر التفاعل بين جزيء الدواء وكل الكيانات الأخرى في النظام البيولوجي على اتباع قواعد محدد. شبكة التفاعلات الدوائية-الدوائية لها دور كبير في إستخدام الأدوية الموجودة حالياً في علاج أمراض لم يكن من المألوف أستخدامها من قبل وذلك عن طريق الكشف عن الخصائص الغير تقليدية للدواء باستخدام تقنيات الذكاء الإصطناعي يتم عرض التنبؤ بخصائص الأدوية كمساهمة هذا البحث في إعادة إكتشاف أستخدامات جديدة للأدوية.

يتم إنشاء مصفوفة ثنائية تعتمد على مجموعة بيانات شبكة التفاعلات الدوائية-الدوائية لوصف كل دواء. في هذه المصفوفة يتم تمثيل كل دواء بواسطة ناقل السمات من جميع الأدوية الأخرى. تم تصميم نموذج التنبؤ باستخدام خوارزميات التعلم الآلي المتعددة ليكون قادراً علي استنتاج خصائصالأدوية بالإضافة إلى التنبؤ بتفاعلاتها الجديدة مع بعضها البعض.

تقدم الدراسة توقعات ل 38 مثالا من العقاقير لديه خصائص مضادة للسرطان. تم التحقق من صحة النتائج عن طريق بحث واسع النطاق لدراسات إكلينيكية سابقة تؤكد هذه التوقعات.

Acknowledgment

First and foremost, Thank God Almighty, Lord of the Worlds. I am thankful to the Informatics program at British University in Dubai (BUiD) for giving me the opportunity to pursue my master degree studies. My sincere compliments to my supervisor Prof Dr. Khaled Shaalan and instructors for teaching me the principles of Computer Science and for guiding me to join the multidisciplinary field of bioinformatics. I am also thankful to my colleagues for their fruitful discussions and feedback.

Finally, I am indebted to my parents for developing my skills in order to understand the true value of education. I would like to thank my family for their support and their blessings throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Nomenclature

ADRs	Adverse Drug Reactions
AI	Artificial Intelligence
AP	Average Position
ATC	Anatomical Therapeutic Chemical (ATC) classification system
AUC	Area Under the Curve
AUPR	Area Under the Precision-Recall curve
AUROC	Area Under the Receiver Operating Characteristic
CTD	Comparative Toxicogenomics Database
DDIs	Drug-Drug Interactions
DNN	Deep Neural Networks
DTIs	Drug-Target Interactions
FAERS	Food and Drug Administration Adverse Event Reporting System
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Study
KEGG	Kyoto Encyclopedia of Genes and Genomes database
LINCS	Library of Integrated Network-based Cellular Signatures
MAP	Mean Average Precision
MeSH	Medical Subject Headings

NSAIDs Nonsteroidal Anti-Inflammatory Drugs

OMIM Online Mendelian Inheritance in Man

PPIN Protein-Protein Interactions Network

RF Random Forest

SIDER Side Effect Resource database

SVM Support Vector Machine

1. Introduction

1.1. Overview

This chapter presents an introductory background about the role of Artificial Intelligence in the drug repurposing (repositioning) process. Drug interactions will be discussed in order to explain the different types of drug interactions and the different databases used by researchers in this field. The problem is explicitly mentioned as well as the rationale and motivation behind this work. This dissertation is designed to answer specific questions related to drug interactions database and its role in drug repurposing. The research aims and objectives are clearly identified and stated. The detailed dissertation structure and design are fully explained. In this chapter, all necessary and related concepts, as well as technical terms, are explained under the Definition Section.

1.2. Background

Machine learning (ML) is a branch of Artificial Intelligence science with a great endeavor to ameliorate the performance of multiple areas of research fields especially the medical field. Classification is one of the cornerstone tasks of ML and Artificial Intelligence (AI) (Weiss & Kulikowski 1991). Classification postulate a predictable class (label) and a reliable collection of attributes to build a high-fidelity dataset. Such dataset is the core of a predictive model to the desired label attribute. The challenge in the classification task is how to find the reliable set of attributes that accurately describe each example in the dataset.

Pharmaceutical drug molecule performs a particular action inside the human body

and interacts with other drugs based on a unique set of pharmacological rules. Factors controlling these rules include chemical structure, gene expression, metabolic pathways, enzymes, carriers, and transporters.

Conventional drug screening approaches are costly and time-consuming as it depends on the experimental extraction of pharmacologically active substances from different possible sources and performs experimental analysis of animals. Then, researchers perform clinical trials on humans to affirm the proper indication with the optimum dose for each drug. Not all drugs succeed to reach the final approval phases on humans due to failure to demonstrate a safe toxicology profile.

As a result, only very few drugs reach the market after a long and expensive process. Indeed, the computational method becomes essential in terms of providing promising drug candidates to bypass the blind screening steps (Chen et al. 2015).

The main DDIs prediction approaches reported by researchers are similarity-based, knowledge-based or mechanism-based methodologies. This thesis combines multiple approaches into one approach that focuses on predicting novel drug indications for already available drugs in the market; this approach is called drug repurposing (repositioning). For instance, (Huang et al. 2018) confirms that antipsychotic drugs have been repositioned for anti-cancer purposes. In addition, the author reviewed the possible mechanisms by which the anti-psychotic medications could possess anti-cancer properties.

In this dissertation, the researcher sheds the light on recent and related efforts by multiple authors who are interested in the same field of study. There are multiple conceptual models behind each author's work and a set of different Machine Learning techniques were applied. Authors predicted new drug-related attributes based on

already established relationships extracted from multiple databases. They validated their findings against extensive literature reviews. Finally, they provided recommendations for drug repurposing.

This study is based on information about drug-drug interactions (DDIs) which is extracted from DrugBank standard database to predict new drug indications using Machine Learning techniques.

1.3. Problem Statement

Novel drug screening before clinical trials lasts for a long time but needs a huge fund. ML provides the guidance for researchers to prioritize their efforts and target the most likely promising drug candidates. This work is designed to investigate the significance of Machine Learning techniques in saving clinical research resources. Previous studies utilized DDIs data combined with other databases for multiple predictions. However, indication-based prediction using DDIs is considered an opportunity for this study to cover.

Introducing new drugs to the market is an ongoing process. Relatively old drugs have a number of reported DDIs much more than that those comparatively new drugs. Safe and effective patient care is much easier to achieve with sufficient drug profile data. However, incomplete DDIs information might lead to serious and unavoidable events.

1.4. Rationale and Motivation

Drug repurposing (repositioning) is the discovery of novel indications for already known drugs. The significance of drug repurposing originates from the economical point of view in terms of saving the money and time spent on new drug discovery steps. The canonical process of drug design and discovery is lengthy and consumes a lot of resources.

Cancer is a prime populace health problem and is one of the major contributors to the worldwide disease burden. The extreme expensive cost of new drug development has led to an increase the concerns towards finding a novel, inexpensive anti-cancer drugs.

The primary idea of introducing ML techniques in such a process is to offer a faster yet an accurate alternative. The finding of this dissertation is a list of top qualified drug candidates to proceed to the next step of the expensive clinical trials.

In addition, application of association rules on DDIs would provide a guidance for rational use of prescribing drug alternatives.

1.5. Research Hypotheses

This dissertation tests the following hypotheses:

- Visual representation and analysis of the Drug-Drug Interactions (DDIs) complex network could function as a classification tool between different drug groups based on their clustering structure.
- Unexplained drug behavior and properties could be explained by graph

analysis of the DDIs network.

- Machine learning classification models could be applied in the Drug-Drug Interactions (DDIs) dataset in order to classify two different drug groups and to predict new drug indications for drug repurposing.

1.6. Aim of Research

This dissertation seeks to provide a fast, inexpensive, reliable, and yet accurate tool for drug repurposing. The study investigates the reliability of DDIs network to infer drug-related properties. The idea is to maximize the role of DDIs data in providing multiple therapeutic options. Furthermore, the study targets the clinical decision support system by providing possible enhancement for better quality of care.

1.7. Dissertation Structure and Organization

In this dissertation, the work is organized in the following order:

Chapter (1) introduces background information about the point of study and explains the significance of ML in drug repositioning. The research problem and motivation are discussed in this chapter. In addition, the aim of work and the hypotheses to be investigated are explicitly mentioned in this chapter. A brief description of the technical terms and domain-specific concepts are clearly stated.

Chapter (2) provides a state-of-the-art literature review of similar studies. In this chapter, the various types of databases used will be mentioned as well as the techniques applied by researchers with an example of all predictable attributes in each article. As

a contribution, a comment on the current limitations is stated with possible solutions for improvements.

Chapter (3) explains the exact steps of the methodology applied in this dissertation. It provides a graphical illustration for clear apprehension. This chapter discusses the dataset acquisition and preparation before visualization and model development. The data mining software and the tool used in network visualization are specifically mentioned. Specific algorithms applied are clearly explained.

Chapter (4) demonstrates the results of three experiments. Each experiment is designed to classify two different drug groups at two steps. First, the results of a classification model are presented in the form of a confusion matrix. Second, the result visually compares each group's clustering pattern in the DDIs graphical network representation.

Chapter (4) provides discussion and interpretation of the reported results. This chapter correlates the results from the visual network analysis with the classifier models. Extensive literature evidence are provided to support the findings.

Chapter (5) summarize the overall study outcome in the conclusion statement. Recommendation for future work is provided as well as the implications for future application.

1.8. Definitions

All study-related terms and concepts are listed and briefly explained to avoid any ambiguity in understanding the research idea. This work is considered a multidisciplinary project. It is expected to combine Computer Science terms with medical and pharmaceutical-related terminologies.

- ML is the application area of Artificial Intelligence (AI) which allows systems to automatically learn as well as to improve its performance based on its own experience without being expressly programmed. ML centers on developing computer programs that can read data and learn from it.

- Pharmaceutical Bioinformatics is a branch of Bioinformatics that focuses on chemical and biological interactions related to drug discovery. However, bioinformatics is a multidisciplinary field of science that combines biological information, mathematics, computer science, and engineering.

- Graph Theory is a visual approach to represent different knowledge domains as a graph for further analysis. The graph is a type of data structure which is extensively used in our real-life. A graph is defined by two elements (nodes, edges). A node or a vertex (N) is a representation of any point or entity in any domain. An edge (E) is a connection between two nodes. The connection could be directed when we have a source and a target. While the undirected graph has no source and target nodes. The edge weight reflects how strong the connection between two different nodes is. In the case of an unweighted graph, it is set to the value (one) as a default.

- Modularity is a measure that distinguishes a network into communities (clusters). High modularity networks have concentrated edges between their nodes inside the same module but sparse edges between their nodes in different modules. Modularity

is used as a community detection technique in the network structures analysis.

- Eigenvector centrality is considered a ranking measure between nodes in the graph theory. Unlike in-degree centrality, a node with high eigenvector centrality score is not necessarily highly connected with a high number of edges. Eigenvector centrality score is more concerned about the node significance than the node degree.

- Drug Indication is the approved use of a particular drug for treating a particular disease.

- Drug repurposing (repositioning) is the process of finding a novel indication for already existing drugs.

- Transcriptomics is the science of studying the transcriptome. The transcriptome is defined as the entire set of RNA transcripts expressed by the genome, under particular conditions or in particular cell lines using methods such as microarray analysis.

- Ribonucleic acid (RNA) is an essential molecule in the biological system. It plays a significant role in the process of gene expression.

- Clustering is a special technique of unsupervised ML that performs grouping of similar data objects. Data points within the same cluster share a similar set of features.

- Classification is a supervised ML technique that assigns a class to a set of data points. It needs a training example and a predictable attribute (Class Label).

- Association Rules are produced by observing data for frequent patterns. The rule is in the form of (if X then Y). The rules are controlled by two parameters (Support and Confidence).

2 Literature Review

2.1 Overview

This chapter represents the main literature review themes. The first theme is a unique collection of top quality peer-reviewed articles related to the dataset type utilized and the applied technique. Authors applied different computational techniques on various drug-related databases. Each article has a unique design and one or more predictable criteria. The main two approaches presented for drug repurposing are (network-based and similarity-based). Summary and conclusion of each article are demonstrated as well as the overall limitations of the current work.

In addition, the second theme in this chapter presents the clinical evidence from the literature that supports anti-cancer predicted properties for each candidate drug.

2.2 Review Articles

Network-based analysis is becoming a highly significant tool to help visualize and understand the connections between drugs and their actions in the body (Berger & Iyengar 2009). Erdos-Renyi (Random or Poisson or Gaussian) networks are considered the basic foundation of the real world complex network development (Newman 2003). Random network assumes that each network is composed of nodes (vertices) which could be individuals or drugs based on the presented domain knowledge. These nodes are connected with each other by connections (edges). The number of edges for each node is called node degree. The degree distribution also known as (neighbor distribution) is considered a characteristic property for each complex network structure.

The principal theory of any random graph is founded on that, each contributing node in the graph has the exact same chances to connect with any other node in the network.

So, the theory expects normal distribution curve by plotting a histogram representing the number of nodes against the number of edges each node has. While this random graph theory could represent some networks in the world, it fails to represent a lot of real networks (Wang & Zeng 2013). The author identified three types of complex network topology (Poisson, Power-Law, and Scale-free). While Poisson topology did not succeed to represent all existing real networks in nature, Scale-free type of networks succeed to represent almost all real networks including DDIs networks. Power-Law (Exponential) is another degree distribution type that represents some networks such as scientists' co-authorship.

This dissertation discusses the concept behind different representation models which were applied to construct distinguishable DDIs networks. Also, it provides suggestions to build a new representation model in order to improve the prediction accuracy of already existing models. Studies presenting the DDIs using networks aim to find common properties describing the quality of each relationship in order to predict unknown interactions between existing drugs or newly discovered drugs using the computational methods rather than experimental ones. The accuracy of the prediction depends on the proper identification of the attributes used to build the network model.

The proposed idea of this work posits the significance of using drug-drug interactions network to extract a set of features for each contributing entity in the network system. The model analyzes a pharmaceutical dataset called DrugBank version (5.1.1) (Law et al. 2013). Each drug unit has a particular structure with multiple attributes. The main proposed idea is to use each contributing member of the DDI network as an attribute for other members. The value for each attribute is binary in nature [0, 1], in which [0] indicates no interaction and [1] indicates the presence of interaction. The constructed matrix consists of rows representing the list of example drugs, and columns representing the features list. 1991 drugs involved in interactions with other drugs as per the DrugBank version 5.1.1 update.

The standard course in the drug development process is costly in terms of money. Time is a significant parameter to be considered as well. In silico methods provide quicker, reliable nevertheless affordable solutions compared to the traditional experimental methods.

Clinical trials on drugs are time-consuming, costly and restricted to a relatively limited number of targets. However, recent studies express that repositioning of already existing drugs can act efficiently as those investigational new drugs.

Earlier studies have confirmed that network analysis is a powerful platform. For example, researchers succeeded to model biological interactions by analyzing biological networks.

This section summarizes related as well as recent work published by researchers focusing on applying machine learning techniques for drug repurposing (repositioning). In general, the main approaches include:

- Application of clustering techniques based on graph theory using complex network visualization and analysis.

- Prediction of a novel drug features or an expected relationship between the drug and other network components based on drug similarity measures. The similarity was calculated from the established connections between the two networks.

There are several biological entities that interact with drugs. For instance, proteins are an essential component of the biological system. Drugs interact with proteins as in the form of target receptors, enzymes.

Gene expression network provides details about gene-related information about drugs as well as diseases. Drugs could have some undesirable side effects, this side effect profile for each drug allow researchers to draw a very informative network to

measure the similarity between drugs. However, researchers face challenges in building prediction models. For example, incomplete or missing reported data lead to a significant decline in the performance measures of any prediction model. Supervised machine learning algorithms are the most sensitive techniques for missing data, the training step for any supervised learning technique necessitates a proper identification of positive data from negative data. Negative data must not be confused by missing data. However, in real life databases, unreported or missing data is considered negative.

(Udrescu et al. 2016) presented a visual analysis to the drug-drug interactions DDIs network. In the first step, the author applied the community detection algorithm in order to identify the modularity-based structure in the DDIs network. Modularity is directly connected to the distribution and density of relationships (edges) between drugs (nodes), which in this study represent DDIs. Nodes in the same modularity were considered to be related because they share the same distribution and density of edges. Each modularity was assigned a distinct color for identification. In the next step, the author applied another algorithm called the topology detection algorithm (Force Atlas II) to discover the main drug clusters within the network structure. Each identified cluster represents a pharmacologically related drug group. However, the distinct color was assigned for each modularity detected in the developed network. The author selected the Force Atlas II layout (Krzywinski et al. 2011) to illustrate the relationship among drugs (nodes) in the network. Force Atlas II is an example of a force-directed layout. It is inspired by the standard laws of physics to visualize networks in the space. In this layout, each node (drug) repulses other nodes as if they have a

similar polarity of charges, while edges (reported interaction between each drug pair) act as a connection springs to attract their nodes. These two conflicting forces of repulsion and attraction eventually reach a balanced state. This final spatial property is expected to provide an explanation and help interpret some findings in the data network. In the Force Atlas II layout, the position of each node depends on other nodes, and the number of connections (edges) each node has. Force Atlas II layout relies on a specific equation to calculate both kinds of forces (attraction and repulsion). The distance (D) between any given nodes (in the geometric space) and the number of edges (E) each node has the main parameters in the equations.

Gephi version (0.9.2) is a specialized software that was used for building and visualizing complex networks (Bastian et al. 2009). It provides the essential analytical tools and measures required to analyze network architecture and discover relations between network components. These measures provide guidance for researchers to interpret unexplained findings.

Ideally, pharmacologically-related drugs should be located close to each other and within the same cluster as well as the same assigned color by modularity. The result shows that clusters of pharmacologically-related drugs are located close to each other. However, the author reports exceptions of particular drugs that showed a tendency to be located next to a different color group of drugs. He concludes that some drugs tend to have unexpected pharmacological actions and act as if they belong to a different group. The author proposed those drugs as promising candidates for drug repurposing according to the predicted new properties.

He validated the results using two versions of DrugBank. The first version (Drug-

Bank 4.1) database to build the graph and predict drug properties and the second version (DrugBank 4.3) to confirm the predicted properties for 85% of the findings. The author argues that complex network analysis provides a high-level of understanding on the

Pharmacological characteristics. Furthermore, the clustering approach can be applied to predict drug-target interactions or customized patient's medicine applications.

Drug-drug interactions were visually analyzed in a previous work done by (Hu & Hayton 2011), the author demonstrated drug-drug interactions (DDIs) network. He affirmed that the presented DDIs network showed scale-free features and followed a power-law frequency distribution. The DDIs network consisted of 966 drugs (nodes) and 3351 interactions (edges). The author selected the top-forty interacting drugs (hot spots). Then, he combined Pharmacokinetics (PK) and Pharmacodynamics (PD) information along with patient demographics to construct a prediction model. The study confirmed the potential activity of a few drugs which was represented as hot spots. The hot spots are connected to a higher number of edges with other drugs compared to the majority of those less active drugs. The findings confirmed that DDIs frequency distribution follows the rules of power-law.

Repositioning drugs for novel indications were investigated by (Hurle et al. 2013). The authors reviewed the in-silico machine learning techniques applied to various drug-related databases. The author highlighted the significant role of transcriptomics data represented by a connectivity map (CMap), side effects (SE) data, and gene-related data represented by the genome-wide association study (GWAS).

In a similar work done by (Setoain et al. 2015), the author utilized transcriptomic data (data related to gene expression) to construct a gene expression signatures dataset. This dataset represented each drug profile as a gene expression signatures data. A similar dataset was constructed for diseases.

The author calculated a drug similarity score to build a model that predicts the relationship between drugs and diseases. After normalization from (0 - 100), he selected the weighted Kolmogorov–Smirnov-like statistic for his analysis. The reported results showed a few cases with prediction above 70% correct.

One of the limitations to this approach is the necessity of having an expertise to domain knowledge personnel for validation. In addition, the cost of computation for the proposed methodology is considered to be high compared to other in silico techniques. In one approach, gene expression signatures could be used alone as a source of information to build a representative drug profile. Another approach is to combine multiple sources of information to build a more accurate drug profile. Example of other sources includes chemical structure database.

(Azuaje 2012) emphasized on the significant contribution of Drug-Target Interactions (DTIs) network as well as Drug-Drug interactions (DDIs) network in the drug discovery process. Analysis of both DTIs and DDIs networks provides a remarkable enhancement in clinical practice. The author provided an example of improving the quality of care for patients with cardiovascular diseases in particular.

(Brown et al. 2016) developed a model to predict new drug therapies. The author combined information from two databases. One database contains gene interactions data, the other database includes chemical exposure drug data. The model successfully predicted five prostate cancer therapies from over 7000 compounds.

(Keane et al. 2015) proposed a model to predict drug targets based on PPIN. The author reported a case of Parkinson’s disease, the model successfully identified 4 proteins related to neurological toxicity. The logic behind the proposed model is

that, if the drug (D1) interacts with protein (P1) and (D2) interacts with protein (P2) then, (P1) and (P2) are similar as long as (D1) and (D2) are similar. In another word, similar drugs interact with similar proteins. Drug interactions network was used as a similarity measure between drugs to predict protein targets (Mei et al. 2012).

In a related article by (Peng et al. 2015), the author applied chemical similarity measures between drugs in order to provide the most promising drug prospect that can fit on specific protein sites. The author assumed that drugs with similar chemical structure will fit on the same protein receptors. This prediction model showed a competitive performance measure (AUC = 0.94).

In a similar work by (Caniza, Galeano & Paccanaro 2017), the author utilized the chemical similarity alone in one model and compared it with a combination between the chemical similarity and the ontology annotations in another model. Prediction models were designed to predict novel drug targets. The assumption in this study was that drugs with similar chemical structures interact with similar targets. Results in both models were AUC 0.59 and AUC 0.69 respectively. The findings supported the significant role of the ontology annotation in enhancing model performance when combined with the chemical similarity. However, the overall model performance is not satisfactory compared to other models performance.

In this paper by (Jin et al. 2017), the author developed a prediction model using the proximal gradient method to solve a regression problem, where each possible DDIs combination between 2 drugs is treated as a task. He used the reported

adverse drug reactions (ADRs) to the FDA as a training dataset for his model. The author claims superior performance over the standard matrix completion methods. The main advantage of this model is the possibility of predicting drug interactions of newly discovered drugs. The author evaluated the effectiveness of his proposed model against a real world database.

According to (Hao, Bryant & Wang 2017), drugs having similar features may enhance the prediction accuracy of their neighbors. In his work, the author integrated network regularization along with the logistic function in order to predict the interactions between drugs and their targets such as (enzymes and receptors). The author highlighted the significance of drug repurposing, he also provided an example of a drug called (Celecoxib) which was initially approved to treat a condition related to bone diseases called (osteoarthritis). However, it was recently approved to be used in the prevention of particular cases of colon cancer.

Gene expression combined with chemical structure data was utilized by (Sawada et al. 2018). The author predicted new drug indications with performance measures reported as (AUC = 0.75).

Similarly, (Raja et al. 2017) predicted adverse drug reactions (ADRs) using drug-gene interactions data. The author selected DDIs features and applied random forest technique as a classifier. He reported F-score of 0.87 as a performance measure of his prediction model.

A semi-supervised model developed by (Peng et al. 2015) to predict drug-protein

interaction using the chemical similarity measures between drugs. The author used (Tanimoto coefficient) as a distance measure in the chemical structure space. He reported competitive results with AUC performance measure equals (0.94) (Hattori et al. 2003).

In (Yoo et al. 2018), the author designed a unique algorithm to predict drug interaction by investigating molecular and phenotypic drug networks. The author introduced a novel method to create an accurate and representative drug profile based on systemic effects data. This profile was used to build a model to predict drug interactions. The author constructed 5,441 profiles of approved and investigational drugs. Those 5,411 drugs are connected to 3,833 phenotypes. He observed a strong possibility of interactions between drug pairs that are highly connected at phenotypes profile. The author reported a successful identification of therapeutic and adverse effects of drugs with high performance measures.

He affirmed the significant role of tracing drug interaction in understanding the mechanism of action in molecular and phenotypic networks. Results reported by the author as scores for therapeutic effect: Area Under the Receiver Operating Characteristic (AUROC) = 0.731 ± 0.021, Area Under the Precision-Recall curve (AUPR) = 0.624 ± 0.003) and adverse effect (AUROC = 0.734 ± 0.033, AUPR = 0.817 ± 0.015) predictions. In addition, when the genetic information associated with the phenotypes was sufficient, the author were able to predict therapeutic (AUROC = 0.731 ± 0.021, AUPR = 0.624 ± 0.003) and adverse effects (AUROC = 0.734 ± 0.033, AUPR = 0.817 ± 0.015) with higher performance.

Drug properties based on Lipinski's rule of five can be utilized in drug repurposing and new DDIs discoveries (Munir, Elahi & Masood 2018). The author collected data related to 2062 drugs, he selected only 1052 of those drugs which fulfill Lipinski's rule of five. The author applied clustering technique and reported 12 clusters formation. He developed DDIs in each cluster using the chemical structure as a measure of similarity between drugs. K-means was applied as a clustering algorithm.

(Liu et al. 2016) extracted DDIs features from DrugBank and combined features related to target pathways from the KEGG database. The proposed model by the author was designed to predict new drug combinations and to avoid potential ADRs. In order to validate the performance accuracy of his model, he argued that DDIs were associated with the predicted pathways.

In a recent work done by (Zhao & So 2018), the author conceded the significant role of machine learning in enhancing the drug repurposing process. He proposed a model to predict drug indication as an output utilizing drug expression as an input data that acquire transcriptomic information from a specific three cell lines (HL60, PC3, MCF7) were treated with a particular drug. The authors selected two groups of drugs called antipsychotics and antidepressants based on the Anatomical Therapeutic Chemical (ATC) classification system. The author compared the performance of more than one machine learning technique including deep neural networks (DNN), support vector machine (SVM), random forest (RF) in order to predict drug indications with binary classifier models. The author used nested three-fold cross-validation to evaluate the model performance

in order to avoid any biased estimation in prediction accuracy. External validation of the reported results was done by comparing the drug list included in the published clinical trials of the predicted indications to the drug list provided by the model. The author argues that, the top drug candidates as a psychiatric drug were considered in ongoing clinical trials, and that many other top hits were verified by already existing studies.

(Zhang et al. 2018) applied a machine learning technique on data related to patients with coronary heart disease. The author provided a set of recommendations to improve the medical practice in terms of prevention and diagnosis for treating this group of patients based on a specific ML algorithm called association rules. In this work, the author proved that ML could provide a high-quality patient care based on a significant decision support system.

(Liu et al. 2013) suggested that DDIs information alone could be utilized to predict physiological properties of drugs. The predicted features could be of much significant value to recommend a new indication for already existing drugs (drug repositioning).

In a previous work performed by (Zhang & Huan 2010), The author proposed ten topological properties extracted from protein interaction networks. Three main groups of protein were identified: drug targets, genes related to disease, and essential genes. The author applied support vector machine and K-nearest neighbor to predict drug-protein targets based on topological properties. The author reported 80% prediction accuracy using 10-fold cross-validation. The

author mentioned the use of logistic regression as a technique for feature selection. However, most similar targets were identified using the k-nearest neighbor method.

Application of network modularity was used by (Yu et al. 2016) to predict novel drug indications for already known drugs (drug repositioning). The author proposed an approach in which he considered the gene relations between drugs and diseases taking the network modularity architecture in his consideration. In his work, the author constructed two networks. The first network represented the relationship between drugs and their side effects. The second network represented the relationship between diseases and their symptoms. The principal logic behind the author's work is the assumption that, similar drugs entail similar diseases. The author identified the distinct cluster modules in both (drugs and diseases) networks. Then, he connected all possible pairs of drugs and diseases. The model was based on known the associations between drugs and diseases, these drugs-diseases relationships were extracted from the Comparative Toxicogenomics Database (CTD) database. One more factor the model relied on, is the local connectivity of each module in both (drugs and diseases) networks. The author predicted potential drug-disease relationships. He validated the results by extensive literature surveys and CTD database. A network connecting drugs with their corresponding side effects was constructed based on a similarity score between drugs. The cosine similarity was used to measure the distance between each pair of drugs. The cosine similarity values start from [0 to 1], where [0] means no common side effects are shared between the selected drug pair.

However, cosine similarity of value equals [1] means exactly the same side effects profile in both drug profiles.

More concerns about drug repurposing in general and drug repurposing for cancer treatment in particular. Researchers reported successful cases of proven activity of drugs against cancer cell proliferation. For instance, tricyclic anti-depressants are a group of drugs that were primarily indicated to treat psychiatric conditions such as depression mood. However, (Cardelli et al. 2018) affirmed that tricyclic anti-depressant-like drugs have the potential to be repositioned for cancer treatment. The discovered properties of such group of medicines provide more options for treatment combinations. Similarly, Nicardipine and Sulindac showed an anti-neoplastic effect against the proliferation of lung cancer cells as reported by (Shi & Zhijian 2018a) (Shi & Zhijian 2018b).

(Hanusova et al. 2015) emphasized on the significance of repurposing drugs for cancer treatment. He mentioned a list of drugs showing a high potential activity against cancer cell proliferation. The list included a drug called Mebendazole, one of the predicted drugs by the proposed DDIs model with remarkable confidence.

In a previous study by (Jamal et al. 2017), the authors investigated the neurological ADRs and have combined multiple drug features including the biological properties such as (enzymes, transporters, and targets). The chemical properties for each drug were also collected as the substructure fingerprints. In

addition, the phenotypic drug properties such as side effects (SE) and therapeutic indications were added to the set of features. In this study, the authors applied a feature selection technique called (relief-based) to identify the most relevant drug properties to the desired predictable target feature which is in this case (ADRs). The authors applied machine learning techniques to build a model in order to predict the neurological adverse drug reactions before the clinical trials testing begins on humans. In addition, in order to demonstrate the efficiency and relevancy of the models, the authors applied to model on anti-Alzheimer drugs to predict their adverse drug reactions. The side effects data were extracted from an open resource (SIDER) database. The models that were based on chemical properties showed accuracy 93.20%. While those models relied on phenotypic properties demonstrated accuracy equals 92.41%. Finally, biological properties succeeded to predict neurological adverse drug reactions with an overall accuracy of 82.11%. However, the authors confirmed that biological-based model was more informative than chemical and phenotypic based models. The authors reported an accuracy enhancement up to 94.18% in model performance due to combining all of the three properties (biological, chemical and phenotypic). Moreover, to prove the predictive ability and to validate the accuracy of the developed models, the models were tested on anti-Alzheimer drugs and on drugs without side effect information recorded in the database (SIDER). The authors believed that the proposed models were highly accurate as well as highly predictive. The authors extracted only those approved drugs from the DrugBank database. They mapped 1991 drugs from DrugBank to the side effect database

(SIDER) using the common and unique drug identifier between the two databases (PubChem CIDs). All the related side-effects and therapeutic indications were accordingly acquired. The constructed dataset included 933 drugs, 5462 side effects, and 3046 therapeutic indications. In this dataset, 933 drugs represented the example set. While all the 5462 side effects and 3046 therapeutic indications represented the feature set. The predictable attribute was the neurological adverse drug reactions. Each drug was represented as a binary matrix of value [0, 1] encoding the absence or presence of each of the corresponding features.

In a dataset studied by (Yamanishi et al. 2008), the author utilized the chemical and the genetic drug information network to predict Drug-Target Interactions (DTIs). The dataset is considered the gold standard for later research work. The author predicted drug targets including enzymes, proteins, and receptors.

Precise identification of drug-target interactions (DTIs) cases in the dataset is essential for machine learning models to generate accurate predictions especially in the field of drug repositioning. In the work done by (Peng et al. 2017), the authors mentioned that only cases experimentally approved as positive cases registered in DTIs databases. Whereas, all undiscovered or even unreported positive cases considered experimentally validated negative values. This problem has a significant effect on model prediction accuracy. The authors proposed a method to screen strongly reported samples of negative drug-target interactions cases in a deposited DTIs database. He calculated the probabilities for all negative samples for being true negatives or true positives. Then he applied support vector machine-based model for optimization. The authors tested the effectiveness of the

proposed method on four different classes of DTIs datasets. They validated the predicted results against unbiased drug database and extensive literature reviews. The authors emphasized that the reported AUC for the proposed method was the highest compared to 6 other states-of-the-art techniques.

The application of computational techniques in the prediction of DTIs has become an essential step in the drug repositioning process. In a recent work by (Luo et al. 2017), the author concluded that the integration between multiple drug-related networks could significantly improve the prediction performance compared to any individual single networks.

Table 2.1Review Summary Review Summary showing machine learning information in the related work. In this table, the key database used in each study is listed as well as the predicted feature.

Study	Key Database	Selected Features	Predicted Features & Comments
(Udrescu et al., 2016)	DrugBank ID (Drug-Drug Interactions)	Bipartite network drug interaction	Pharmacological properties and drug behavior
(Setoain et al., 2015)	DrugBank,OMIM, KEGG and PGDB	Gene expression Network	Prediction > 70% drugs-diseases relationships
(Brown et al., 2016)	CTD, GEO	Gene Network Plus Chemical attributes	Predicted 5 Prostate Cancer Therapies
(Keane et al., 2015)	KEGG, OMIM, and PPIN	The bipartite network between drugs and proteins	Predict protein targets for drugs
(Peng et al., 2015)	DrugBank, KEGG, ChEMBL, Matador	Chemical attributes and ADRs	Predict protein targets for drugs. AUC = 0.94
(Caniza et al., 2017)	DrugBank, MeSH	Chemical attributes and Ontology annotations	Predict drug targets. AUC = 0.59-0.69
(Jin et al., 2017)	ADR, DDIs, FAERS	Adverse drug reactions network (ADRs)	Predict DDIs, ADRs
(Sawada et al., 2018)	LINCS, ChEMBL	Gene expression combined with chemical structure data	Predict drug indications. AUC = 0.75
(Raja et al., 2017)	DrugBank, CTD	Gene interactions and DDIs	ADRs. F-score = 0.87
(Yoo et al., 2018)	DrugBank, MeSH, OMIM, CTD	Molecular and Phenotypic drug networks	Predict Therapeutic and ADRs, AUROC = 0.731, AUPR = 0.817
(Munir et al., 2018)	ZINC	Chemical attributes	Predict DDIs
(Liu et al., 2016)	KEGG	DDIs, Target Pathways	ADRs
(Zhao & So, 2018)	CTD	Gene expression	Predict drug indications
(Zhang & Huan, 2010)	PPIN	Protein interactions network	Predict protein targets for drugs. 80% prediction accuracy
(Yu et al., 2016)	CTD	Gene-drug, Gene-disease	Predict drug indications
(Jamal et al., 2017)	DrugBank, SIDER, PubChem CIDs	Biological, chemical, phenotypic	Predict neurological ADRs

Table 2.1Review Summary Review Summary showing machine learning information in the related work

2.3 Clinical Support of Predicted Properties.

In 2012, (Fond, G and Macgregor, A and Attal, J and Larue, A and Brittner, M and Ducasse, D and Capdevielle, D 2012) reviewed 93 studies related to the effect of antipsychotic drugs on cancer development. The author expressed his ideas and concerns towards the possibility of antipsychotic drugs to function as anti-cancer drugs.

(Singh & Sharma 2018) confirmed the cytotoxic properties of two substances called (berberine and sanguinarine) alkaloids. The authors reported a decrease in the activity of an enzyme called benzphetamine n-demethylase as a result of using berberine and sanguinarine alkaloids. These findings do not have a direct mechanism to explain the role of benzphetamine activity against cancer cells. However, it indirectly correlates benzphetamine level to the anti-cancer properties of the tested alkaloid substances.

(Yin et al. 2018) mentioned the genetic correlation between alosetron and bladder cancer.

(Bruno et al. 2018) acknowledge the preventive role of anti-thrombotic agents in cancer cell development as well as their role in reducing metastatic infiltration and overall mortality.

Suramin used in the treatment for the African type of trypanosomiasis. However, it is considered as an investigational drug, clinical trials are currently testing anti-cancer properties. (Su et al. 2018)argues that the mechanism of action of suramin is not clear, and it might be due to its inhibitory effect on the DNA. Nevertheless, he supports future research to investigate the anti-cancer

properties of suramin.

Tetracycline and doxycycline belong to Tetracycline's drug group. (Lokeshwar, Escatel & Zhu 2001) affirmed the activity of doxycycline as a significant anti-cancer cell agent. In addition, tetracycline was investigated in clinical trials for its anti-cancer properties against lung and breast cancer.

Geldanamycin shows activity against colorectal and pancreatic cancers as reported by (Mayor-López et al. 2014) (Mohammadian et al. 2017). One reported clinical trial investigated the effectiveness of geldanamycin against hematological malignancy.

Mebendazole belongs to anthelmintic drug group. It was approved for the treatment of different types of worms. However, (Rubin et al. 2018) supported the anti-cancer properties of Mebendazole. The author said that Mebendazole potentiated the immune system in the body which could be related to its anti-cancer activity.

In 1963, ethyl carbamate (urethan / urethane) was removed from the markets in Canada, USA, and the UK due to carcinogenic effects. However, in 2018, (Soni & Soman 2018) investigated the anti-cancer properties of aminocoumarin derivatives. One of the screened derivatives included ethyl carbamate moiety.

Sirolimus belongs to a group of drugs called macrolides. It has potent immunosuppressant activities and its primary indication is the prophylaxis against organ rejection. Sirolimus is an investigational drug for bladder cancer treatment in the current clinical trials pipeline. In a recent work by (Jung et al.

2017), the author confirmed the anti-cancer properties of sirolimus.

A recent review by (Lian et al. 2018) concluded that all published work by researchers who investigated the anti-cancer properties of gemfibrozil provided evidence of its activity against cancer cells in human.

(Garcia-Quiroz & Camacho 2011) argued that histamine supports cell proliferation of both normal as well as cancer types. In addition, the role of histamine and inflammation in cancer progression was confirmed by (Coussens & Werb 2002). However, the reason why histamine was classified as anti-cancer might be due to the fact that some anti-cancer drugs have carcinogenic activity as well as reported in (Lien, E. J. and Ou, Xing-chang 1985).

An invention registered by (Shi & Zhijian 2018c) demonstrated the efficacy of Desogestrel against colon cancer as well as breast cancer cells. These findings confirm the predicted results by the proposed computational model in this thesis.

As mentioned in (Hanusova et al. 2015), metformin and pioglitazone primarily indicated for the treatment of diabetes (a disease characterized by high blood glucose level). However, clinical experimentation showed a high potential activity against cancer cells. These findings support the idea of repositioning metformin and pioglitazone for cancer treatment. Albendazole and mebendazole are another examples provided by the authors as a suggestion for drug repositioning. Albendazole and mebendazole belong to a drug group called (anthelmintic) which primarily indicated for the eradication of worm infection. Nevertheless, clinical evidence confirmed the anti-cancer properties for

Albendazole as well as Mebendazole. This clinical evidence confirms the predicted properties of the proposed model.

In a previous study done by (Cohen, Dembling & Schorling 2002), the authors observed lower rates of cancer cases in patients with schizophrenia. The authors believe that antipsychotic drugs might possess anti-cancer properties. Clinical and computational studies proved the anti-cancer activity of members belong to this group. For instance, Phenothiazine inhibited the proliferation of various cancer cells at concentrations similar to those found in psychotic patients (Nordenberg et al. 1999). In addition, phenothiazine showed anti-cancer activity against glioma cells (Gil-Ad et al. 2004). Moreover, (Qi & Ding 2013) applied computational methods to analyze protein interactions network against the mechanism of action of phenothiazine. The authors confirmed the potential anti-cancer mechanism by phenothiazine. A similar work done by (Yde et al. 2009) mentioned that the anti-cancer effects of Tamoxifen on breast cancer cells enhanced by chlorpromazine when given in combinations.

(Barron et al. 2011) observed reduced rates in both mortality as well as the progression of breast cancer in patients receiving nonspecific β-blocker (propranolol) compared to those patients receiving specific β-blocker (atenolol). The authors concluded that preclinical observations provided associations between the inhibition of β2-adrenergic signaling pathway and the reduced rates in breast cancer progression and mortality.

(Lee et al. 2007) mentioned that chlorpromazine inhibits proliferation of tumor

cells via the inhibition of mitotic kinesin. The authors added that the combination of chlorpromazine and pentamidine resulted in a more potent inhibitory action on cancer cells. This additional benefit could be due to the observed synergistic action between the two drugs on mitotic inhibition.

Fluoxetine is a drug belongs to the anti-depressant class. However, (Peer & Margalit 2006) reported that fluoxetine could potentiate the cancer cell response to anti-cancer drugs via the possible chemo-sensitization mechanism.

The association between metformin and risk reduction cancer-related mortality in diabetic patients had been reviewed and confirmed by (Franciosi et al. 2013). However, the authors emphasized the necessity of performing randomized clinical trials to investigate the efficacy of metformin on cancer cells.

(Hosono et al. 2010) conducted a clinical trial to evaluate metformin activity on cancer cells. The authors administered metformin daily for 1 month to non-diabetic patients with a confirmed cancer diagnosis. The study confirmed a decline in proliferation rates of colorectal epithelial cells. Activity against breast cancer was also associated with metformin according to (Oliveras-Ferraros et al. 2011), (Taylor et al. 2013). These clinical findings confirm the predicted anti-cancer properties of metformin.

Pioglitazone is another drug indicated for the treatment of diabetes. However, animal studies showed activity against liver and lung cancer as mentioned in (Hanusova et al. 2015).

FDA initially approved Leflunomide for treatment of rheumatoid arthritis

disease. However, recent researches recognized it as a potent anti-cancer drug. (Zhang & Chu 2018) discussed the results of leflunomide's effectiveness as well as the possible mechanisms of action of this drug to be an anti-cancer agent.

2.4 Current Limitations

The prediction performance of network-based models is limited compared to knowledge-based and statistical models. The binary matrix properties of the developed DDIs network are not greatly supporting the use of such network-based models. In a binary matrix of values [0, 1], [0] means absence of interactions, and [1] means presence of interactions. In other words, [0] value represents both cases where no interactions are reported and those cases where undiscovered (unreported) interactions occur.

In addition, newly discovered drugs and drugs with limited use (orphan drugs) do not have the average number of reported interactions in their profile compared to relatively old drugs. This leads to inaccurate profile information and negatively affects the prediction model performance.

3 Research Material and Methods

3.1 Overview

The methodology steps in this dissertation are illustrated in (Figure 03.1 Methodology Steps Illustrated). The initial processing step extracts the Drug-Drug Interactions (DDIs) in the form of nodes and edges. Then, it creates a network graph representation. Next, it analyzes the DDIs network to investigate the unique cluster architecture. Afterward, it constructs a binary matrix that describes each drug profile as a set of features. Finally, it builds a model to predict drug indications using the binary matrix and a predictable class as a label attribute based on ATC classification.

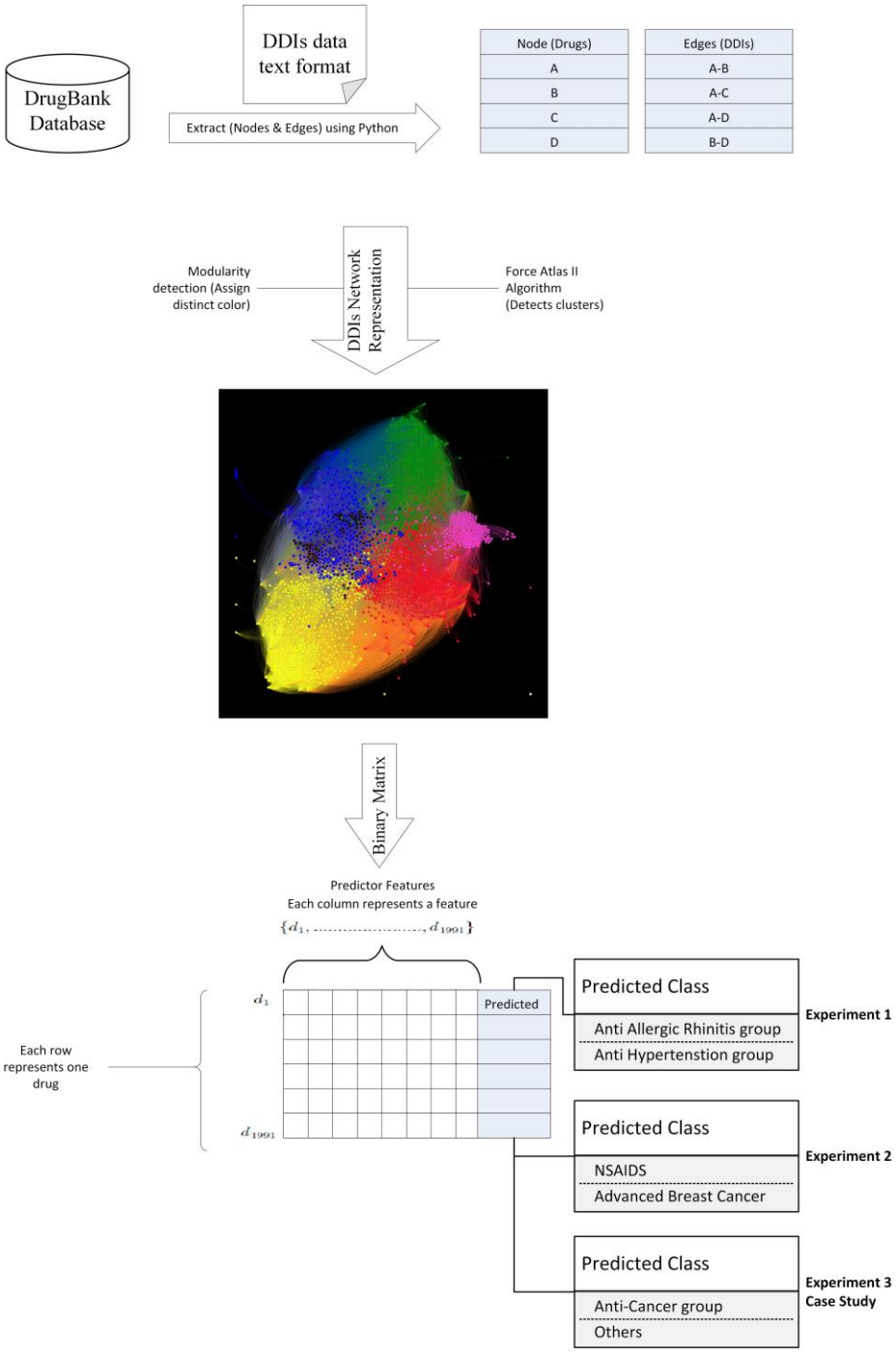

Figure 03.1 Methodology Steps Illustrated

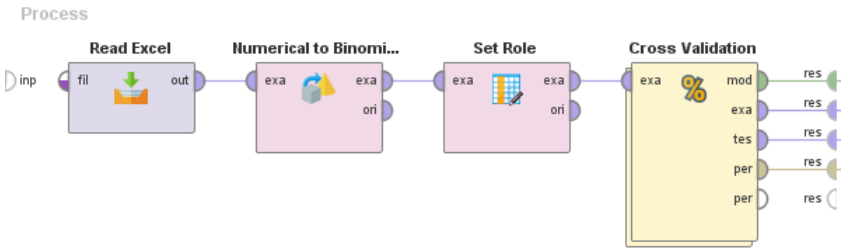

Figure 3.2. Modeling Process

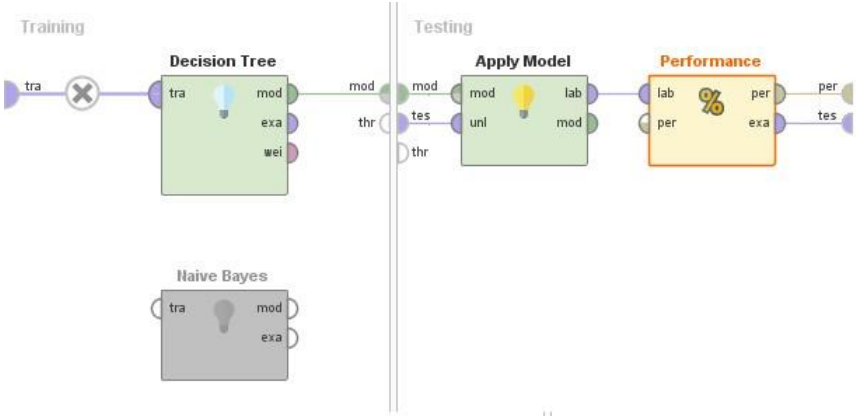

Figure 3.3 Classification Model Training and Testing Steps

Figure 3.2. Modeling Process), Figure 3.3 Classification Model Training and Testing Steps) represents the classifications model steps. Training and testing steps are sub processes included inside the cross validation block. DT and NB are applied each one at a time.

3.2 Data Acquisition and Dataset Construction

DrugBank is an open source database contains information related to drugs and describes its features. DrugBank lists drug interactions as a text format and it has regular updates. DrugBank (Law et al. 2013) Database Version 5.1.1 — recently released on July 3rd, 2018 was used in this work. Drug interactions are processed using Python (programming language) to extract drugs as nodes and the connection between each interacting drug pair as an edge. In this dissertation, 1991 unique drugs were identified as nodes. Each node has at least one interaction reported with another drug node. A total of 210,850 DDIs are reported as edges between nodes.

The first step in this work is a knowledge representation in the form of a network visualization to illustrate the connections between interacting drugs. Drug clusters were identified by running the modularity-based algorithm and Force Atlas II layout algorithm.

The Modularity-based algorithm in our work identifies distinct communities (clusters) within a network. Each community has concentrated edges between their nodes. In other words, the nodes within any module tend to have more edges between themselves than edges with other nodes in different modules. Modularity analysis is used as an indicator of community detection in a network. Each modularity detected were identified by distinct color. Gephi v (0.9.2) (Bastian et al. 2009) is a visualization tool based on the Java platform.

The Force Atlas II layout algorithm in our work is used to allow nodes to be in close position to all nodes they have a connection with. In addition, it allows nodes without edges to be far away from each other.

The second step in dataset construction is to build a binary matrix from nodes and edges. The interactions network between drugs is extracted from the database in a text format. A list of each pair of interacting drugs is extracted using Python. Drug profile binary matrix with dimensions (1991*1991) is constructed in which rows represent a list of drugs, columns represent a list of features.

The list of features are those drugs involved in interactions with the example set. Where, the binary value (v = [0, 1]) indicates whether or not there is an interaction. Decision Tree and Naive Bayes classifiers were applied using (RapidMiner Studio Version 9.0). The 10-folds cross-validation evaluation technique was used to validate the results. In this thesis, we performed 3 experiments. In each experiment, two different drug groups were compared at a time.

3.2.1 Experiment (1): Anti-hypertensive drugs VS Anti Allergic rhinitis drugs

In this experiment, two groups of drugs were selected based on the highest number of population affected and the maximum number of DDIs their members have with other drugs. The first group is listed in Table (Table 03.1 List of Anti-Hypertensive drugs), this group includes 69 drugs and is indicated for the treatment of hypertension disease. The second group includes 48 drugs and is listed in Table (Table 3.2 List of Allergic Rhinitis drugs).

First, the study visualizes the node location of the two groups under investigation on the DDIs network graph using Gephi v (0.9.2). Graph analysis is performed to calculate the number of clusters and node degree.

Then, the DDIs data is used to build a binary matrix and apply classification algorithms (Decision Tree and Naive Bayes) to confirm the visual graph analysis with a measurable performance measure. The binary matrix includes 1991 feature attributes, two class labels, and a total of 117 example set. RapidMiner Studio V 9.0 is utilized to build, train, and validate the classification model. RapidMiner is a leading data mining tool. It is a graphical user interface based on Java platform.

No.	Drug Name	No.	Drug Name	No.	Drug Name
1	Acebutolol	31	Isradipine	61	Telmisartan
2	Amiloride	32	Labetalol	62	Terazosin
3	Amlodipine	33	Lisinopril	63	Timolol
4	Atenolol	34	Losartan	64	Torasemide
5	Bendroflumethiazide	35	Methyclothiazide	65	Trandolapril
6	Betaxolol	36	Methyldopa	66	Triamterene
7	Bisoprolol	37	Metolazone	67	Trichlormethiazide
8	Captopril	38	Metoprolol	68	Valsartan
9	Carteolol	39	Minoxidil	69	Verapamil
10	Carvedilol	40	Moexipril		
11	Chlorothiazide	41	Nadolol		
12	Chlorthalidone	42	Nebivolol		
13	Cilazapril	43	Nicardipine		
14	Clonidine	44	Nifedipine		
15	Diltiazem	45	Nisoldipine		
16	Doxazosin	46	Nitroprusside		
17	Enalapril	47	Olmesartan		
18	Eplerenone	48	Oxprenolol		
19	Eprosartan	49	Pargyline		
20	Felodipine	50	Penbutolol		
21	Fosinopril	51	Perindopril		
22	Furosemide	52	Pindolol		
23	Guanabenz	53	Polythiazide		
24	Guanethidine	54	Prazosin		
25	Guanfacine	55	Propranolol		
26	Hydralazine	56	Quinapril		
27	Hydrochlorothiazide	57	Ramipril		
28	Hydroflumethiazide	58	Reserpine		
29	Indapamide	59	Spirapril		
30	Irbesartan	60	Spironolactone		

Table 03.1 List of Anti-Hypertensive drugs

No.	Drug Name	No.	Drug Name
1	Acetaminophen	25	Fexofenadine
2	Alimemazine	26	Flunisolide
3	Astemizole	27	Fluticasone propionate
4	Azatadine	28	Hydrocodone
5	Azelastine	29	Hydrocortisone
6	Betamethasone	30	Hydroxyzine
7	Brompheniramine	31	Loratadine
8	Budesonide	32	Methdilazine
9	Caffeine	33	Methylprednisolone
10	Carbinoxamine	34	Methylscopolamine bromide
11	Chlorphenamine	35	Montelukast
12	Ciclesonide	36	Olopatadine
13	Clemastine	37	Phenindamine
14	Codeine	38	Pheniramine
15	Cortisone acetate	39	Phenylephrine
16	Cyclizine	40	Phenylpropanolamine
17	Cyproheptadine	41	Prednisolone
18	Desloratadine	42	Prednisone
19	Dexamethasone	43	Promazine
20	Dexbrompheniramine	44	Pseudoephedrine
21	Dextromethorphan	45	Scopolamine
22	Diphenylpyraline	46	Triamcinolone
23	Doxylamine	47	Tripelennamine
24	Ephedrine	48	Triprolidine

Table 3.2 List of Allergic Rhinitis drugs

3.2.2 **Experiment (2): Advanced breast cancer drugs VS NSAIDs**

Similarly, this experiment compares another two different groups of drugs. The first group is listed in Table (Table 3.0.3 List of Advanced Breast Cancer drugs investigated.), this group contains 25 drugs which are indicated for the treatment of advanced stage breast cancer disease. The second group includes 40 drugs, which belongs to a drug group labeled as (non-steroidal anti-inflammatory drugs), listed in Table (Table 3.0.4 List of NSAID drugs investigated.).

The study followed the exact steps followed in Experiment (1). First visualization of nodes position in the DDIs network graph. Then, apply classification technique on the DDIs binary matrix using the corresponding predictable class attributes of Experiment (2).

No.	Drug Name
1	Paclitaxel
2	Trastuzumab
3	Afatinib
4	Anastrozole
5	Goserelin
6	Capecitabine
7	Fulvestrant
8	Megestrol acetate
9	Cisplatin
10	Enzalutamide
11	Gemcitabine
12	Lapatinib
13	Trastuzumab emtansine
14	Vandetanib
15	Cediranib
16	Dasatinib
17	Decitabine
18	Docetaxel
19	Everolimus
20	Letrozole
21	Vorinostat
22	Bosutinib
23	Exemestane
24	Eribulin
25	Sunitinib

Table 3.0.3 List of Advanced Breast Cancer drugs investigated.

No.	Drug Name	No.	Drug Name
1	Aceclofenac	21	Fenbufen
2	Flurbiprofen	22	Kebuzone
3	Mefenamic acid	23	Meclofenamic acid
4	Nabumetone	24	Benoxaprofen
5	Suprofen	25	Naproxen
6	Diclofenac	26	Parecoxib
7	Misoprostol	27	Zomepirac
8	Oxyphenbutazone	28	Etoricoxib
9	Phenylbutazone	29	Ibuproxam
10	Lumiracoxib	30	Indoprofen
11	Oxaprozin	31	Niflumic Acid
12	Tenoxicam	32	Valdecoxib
13	Indomethacin	33	Etodolac
14	Ketorolac	34	Ketoprofen
15	Lornoxicam	35	Nimesulide
16	Piroxicam	36	Tolmetin
17	Azapropazone	37	Fenoprofen
18	Meloxicam	38	Ibuprofen
19	Tiaprofenic acid	39	Rofecoxib
20	Celecoxib	40	Sulindac

Table 3.0.4 List of NSAID drugs investigated.

3.2.3 **Case Study: Anti-Cancer Drug Prediction.**

This case study seeks to draw the classification boundary between the groups of drugs showing anti-cancer properties against all other drugs using DDIs binary matrix dataset. The experiment selected 250 drugs listed in DrugBank v 5.1.1 and is classified as an anti-cancer group according to the ATC system. All remaining drugs in the DrugBank are labeled as a test group, this group contains 1741 drugs (1991-250).

In this case study, we evaluate the classification performance of the model as well as investigate the model predictions for repurposing. In another word, we need to identify the drugs having the potential to be classified as an anti-cancer group.

3.3 Prediction Algorithms

3.3.1 Decision Tree (DT)

A decision tree is a known classification technique. It splits the data into smaller subsets based on certain criteria. The end result is in the form of a tree representation with its root node at the top while decision nodes have one or more arms based on the split feature. The terminal part is called the leaf node which represents the class decision. ID3 (Quinlan 1986) is the most common algorithm that we used to build the decision tree based on two parameters (entropy and information gain).

Entropy. The ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy equals to zero. However, entropy equals one whenever the sample is equally divided.

Information gain. The information gain is dependent on the decline in entropy value after each split on the dataset over an attribute. Building a decision tree is basically concerned with identifying the split attribute that returns the maximum information gain score. In other words, the most consistent branches are associated with information gain value.

3.3.2 Naive Bayes (NB)

Naive Bayes is one of the supervised ML algorithms. It is based on a basic assumption in which the features are independent of each other. Each feature is assumed to have an independent distribution. Therefore, the covariance in features is not considered as a factor that might affect the performance.

3.3.3 Deep Learning (DL)

Recently, Deep Learning techniques are designed for scientific domains which store and process big data, such as the area of bioinformatics (Bacciu et al. 2018). Deep learning is basically adding multiple hidden layers in the neural network architecture. Each hidden layer is getting trained for optimum feature selection technique. The number of nodes in each hidden layer is decreasing towards the direction of the output layer (prediction layer).

3.4 Model Performance Evaluation Measures

A total of 3 ML models were generated for DDIs, which were evaluated using multiple statistical measures, such as Accuracy, Precision, Recall, and F-measure.

Accuracy (A) is the ratio of correctly identified examples either positive or negative in relation to the entire example set.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Where: TP = True positive; FP = False positive; TN = True negative; FN = False negative

Precision (P) is the ratio of correctly identified positive examples in relation to all predicted positive examples.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall (R) is the ratio of correctly identified positive examples in relation to all true positive examples. Recall measure could be referred to as (True Positive Rate).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

False Positive Rate (FPR) is the ratio of incorrectly predicted positive examples in relation to all true negative examples. FPR expresses the correctly identified negative examples.

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{12}$$

F-measure is the harmonic mean of precision and recall.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (13)$$

4 Results and Discussion

4.1 Overview

This chapter illustrates the visual positioning of each drug node and their relative distance compared to other nodes of the same class. In addition, it presents the Confusion Matrix for each classifier in Experiment (1), Experiment (2), and the case study.

Each drug group is identified by a distinguishable color to visually spot their location on the constructed DDIs network. The classification model provides a quantifiable measure of the visually drawn boundary between each drug pair. The visual graph analysis provides interpretation of the model prediction candidates for repurposing.

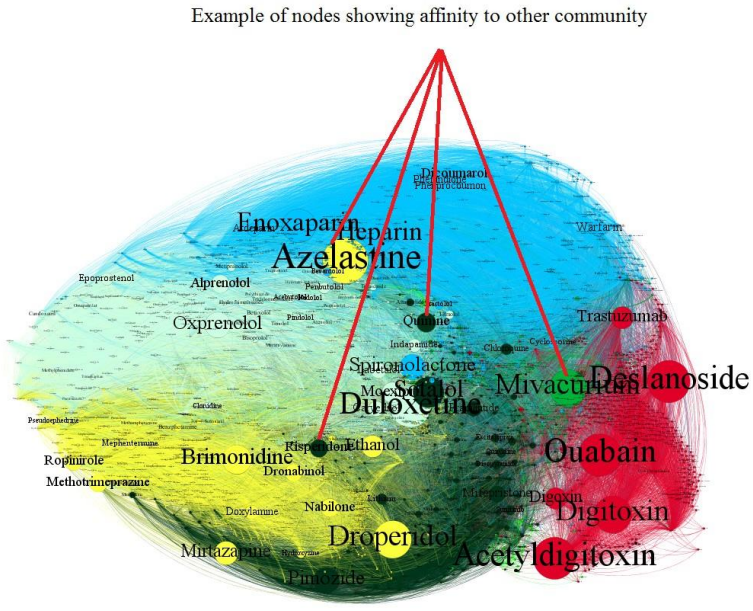

Figure 4.1 Overall DDIs network visualization

The visual representation of the DDIs complex network is illustrated in Figure (Figure 4.1 Overall DDIs network visualization). It represents a number of isolated nodes with distinct colors from their corresponding neighbors. This indicates that the node has been classified as a member of one group but yet, it also shows some affinity to a different group of drugs. For example, Azelastine is initially classified as an antihistamine (a drug used for the treatment of allergic conditions) but its location on the complex network shows affinity to a different community that is characterized by increased bleeding tendency. Clinical practice confirms that nose bleeding is a commonly reported side effect of Azelastine. This finding could provide a clue about the association between the drug node position on the DDIs network graph and its uncovered pharmacological features.

4.1.1 Experiment (1): Anti-Hypertensive Drugs VS Anti Allergic Rhinitis Drugs

Figure 4.2 Anti-Hypertension VS Anti Allergic Rhinitis) shows the node clustering of two drug groups in the DDIs network. Red nodes represent (Anti- Hypertensive group), blue nodes represent (Anti-Allergic Rhinitis group). Red nodes show a relatively distinct boundary compared to the sporadic distribution of blue nodes. The degree distribution for each drug group over each modularity cluster is reported in

Table 4.2 Degree Distribution Experiment (1). It confirms the localization of red nodes in modularity number (0). 61 drugs belong to Anti-Hypertensive Group are clustered in modularity number (0) which contains a total of 301 drugs. However, blue nodes are scattered over 5 modularity clusters. Further analysis of the graph shows 4 blue dots are located in modularity number (0) where the majority of red dots are.

Accuracy: accuracy: 100% +/- 0% (micro average: 100%)			
	True Anti HTN Class	True Allergic rhinitis Class	Class Precision
Predicted Anti HTN Class	69	0	100.00%
Predicted Allergic rhinitis Class	0	48	100.00%
Class Recall	100.00%	100.00%	

Table 4.1 DT Performance Metrics (Anti-Hypertensive VS Anti Allergic rhinitis)

Table (Table 4.1 DT Performance Metrics (Anti-Hypertensive VS Anti Allergic rhinitis)) reports the Confusion Matrix of the Decision Tree (DT) classification algorithm applied over the Anti-Hypertensive and Anti-Allergic Rhinitis group. It shows a perfect classification measures with a reported 100% for all performance indicators (Accuracy, Precision, Recall). The DT model demonstrates the capability of perfectly identifying each member of the two investigated groups based on the DDIs binary matrix features.

Modularity No.	Anti-Allergic Rhinitis Group	Relative %	Anti-Hypertensive Group	Relative %	Other Drugs	Grand Total
0	4	1.33%	61	20.27%	236	301
1	0	0.00%	0	0.00%	222	222
2	1	1.18%	0	0.00%	84	85
3	25	6.83%	3	0.82%	338	366
4	14	2.92%	4	0.84%	461	479
5	5	0.93%	3	0.56%	529	537
6	0	0.00%	0	0.00%	2	2

Table 4.2 Degree Distribution Experiment (1)

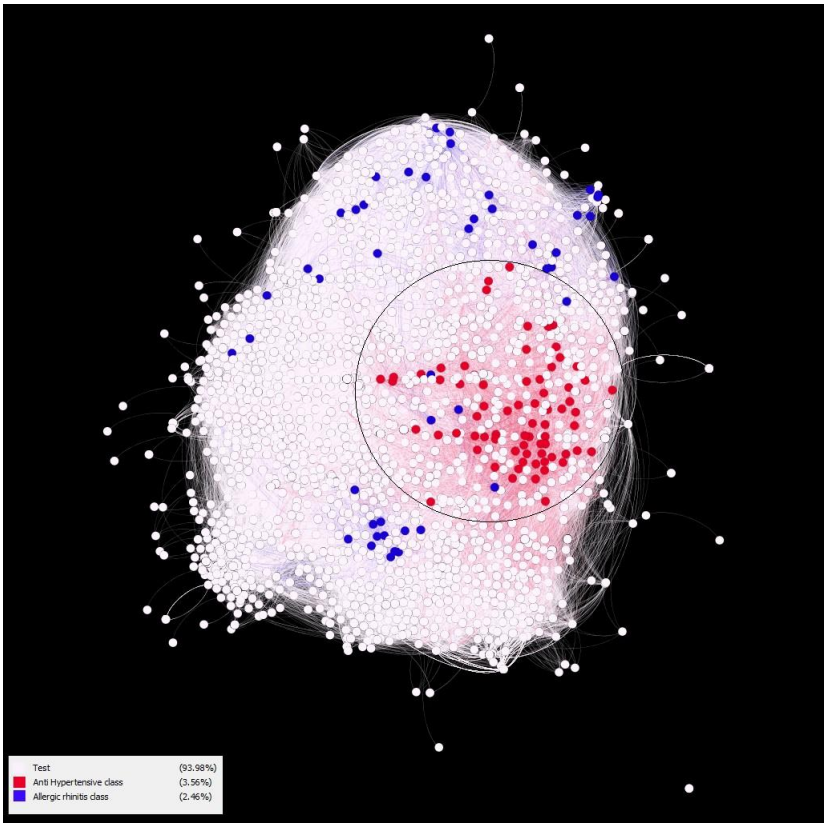

Figure 4.2 Anti-Hypertension VS Anti Allergic Rhinitis

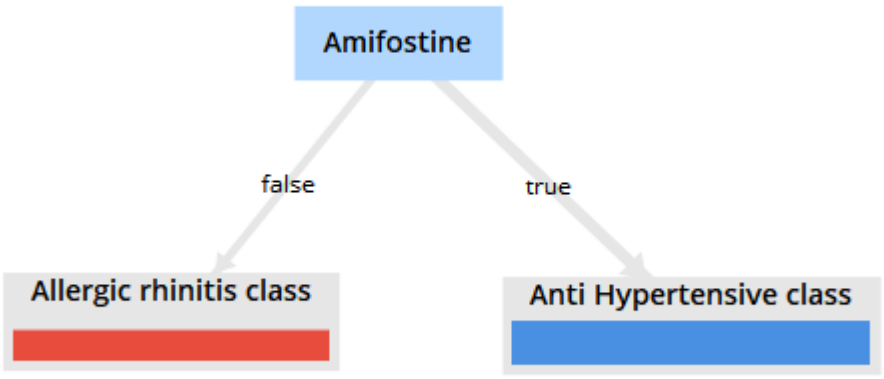

Figure 4.3 Decision Tree Result Experiment (1)

Figure (Figure 4.3 Decision Tree Result Experiment (1)) demonstrates the DT results of Experiment (1). The results summarize the logic of the classification concluded by the DDIs dataset provided. DT concludes that the Amifostine node is a key determiner to be considered in order to differentiate between the Anti-Hypertensive group and Anti-Allergic Rhinitis group. No reported DDIs between members of Anti-Allergic Rhinitis group and Amifostine. However, all members of the Anti-Hypertensive group have a documented DDIs with Amifostine.

Figure (Figure 4.4 Amifostine Drug Interactions Network) depicts the Amifostine edges in the DDIs network graph, where Amifostine is the central node. The graph confirms the presence of connections between Amifostine with all members of the Anti-Hypertensive group (red nodes). However, no connection between Amifostine with any member of the Anti-Allergic Rhinitis group (blue nodes) was observed.

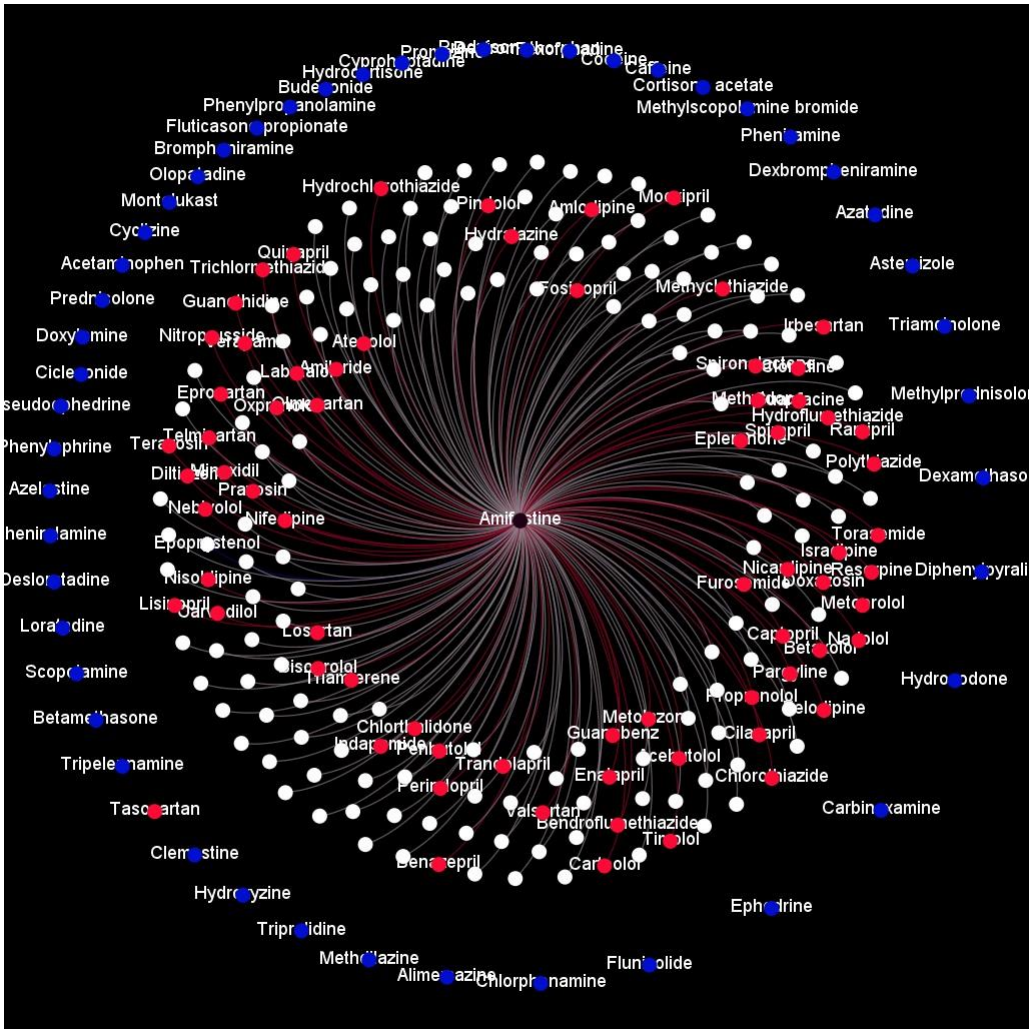

Figure 4.4 Amifostine Drug Interactions Network

Accuracy: 99.17% +/- 2.50% (micro average: 99.15%)			
.	True Anti HTN Class	True Allergic rhinitis Class	Class Precision
Predicted Anti HTN Class	69	1	98.57%
Predicted Allergic rhinitis Class	0	47	100.00%
Class Recall	100.00%	97.92%	

Table 4.3 NB Performance Metrics (Anti-Hypertensive VS Anti Allergic rhinitis)

Table (Table 4.3 NB Performance Metrics (Anti-Hypertensive VS Anti Allergic rhinitis)) presents the performance measures of Naive Bayes (NB) classifier.

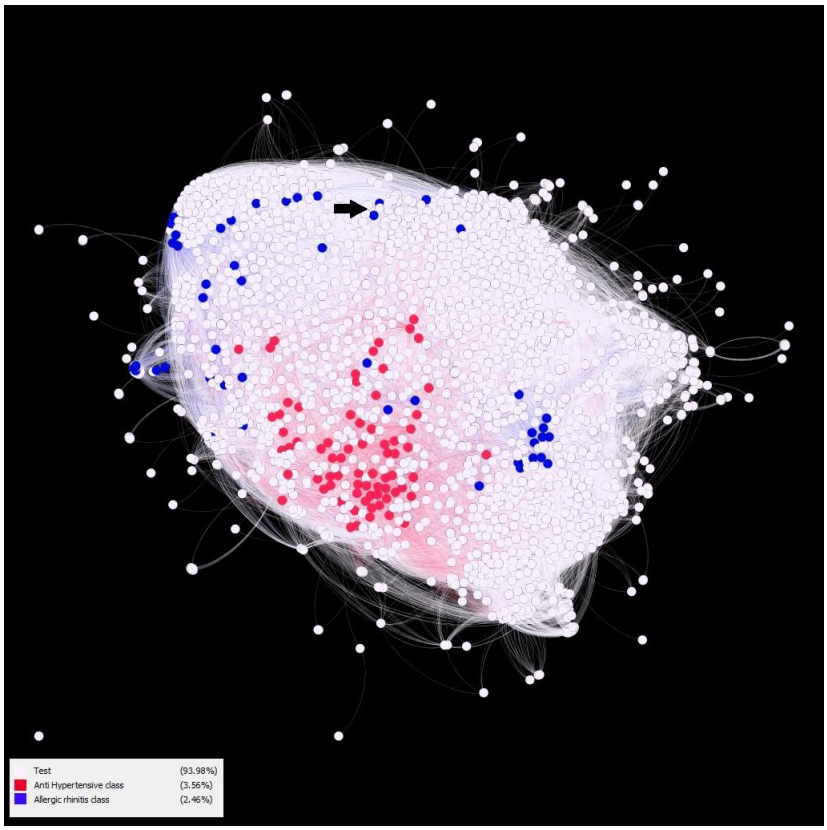

Figure 4.5 Clemastine Position

NB shows 100% Recall of all Anti-Hypertensive group. However, Precision for the same group was reported as 98.57%. One drug (Clemastine) belongs to the Anti-Allergic group was classified as Anti-Hypertensive by NB. The position of Clemastine in the DDIs graph network as illustrated in figure *(*

Figure 4.5 Clemastine Position) do not fully agree with the predicted feature.

Drug Name	Modularity No.	Node Degree	Drug_Group
Verapamil	5	838	Anti-Hypertensive class
Diltiazem	5	807	Anti-Hypertensive class
Isradipine	5	749	Anti-Hypertensive class
Clemastine	5	715	Allergic rhinitis class
Caffeine	5	317	Allergic rhinitis class
Dextromethorphan	5	275	Allergic rhinitis class
Astemizole	5	163	Allergic rhinitis class
Montelukast	5	104	Allergic rhinitis class

Table 4.4 Modularity Number (5) Degree Distribution

Table (Table 4.4 Modularity Number (5) Degree Distribution) displays the drug nodes within modularity number (5). The total nodes are 8, 3 of them belong to an Anti-Hypertensive group, and the remaining 5 belong to the Anti-Allergic group. Clemastine has a node degree value equals 715. This value is significantly higher than the degree of other nodes of the same group.

4.1.2 Experiment (2): NSAIDs VS Advanced Breast Cancer Drugs (ABCD)

The result in this section presents the relative node position of two different drug groups. Figure (

Figure 4.6 Advanced Breast Cancer Drugs VS NSAIDs) illustrates the clustering pattern of the NSAIDs group (blue nodes) and ABCD group (red nodes). The two groups have almost clear defined boundaries without overlapping areas. The nodes of the same group are located relatively close to other nodes of the same group than to nodes of the different group. Unlike all red nodes, Cisplatin shows a tendency towards the area of blue nodes. Table (

Table 4.5 Naive Bayes Performance metrics (NSAIDs VS Advanced Breast Cancer)) reports the performance results of NB classifier, the prediction agrees with the graph visualization analysis. NB successfully identified 24 out of 25 nodes of the ABCD Group (A), 1 node (Misoprostol) was predicted as a member of Group (A), whereas, it belongs to Group (B). Table (Table 4.6 Node Distribution Experiment (2)) displays the degree distribution of all nodes overall modularity clusters. It is obvious that modularity number (4) contains all Group (B) members plus one member of Group (A).

Accuracy: 97.14% +/- 5.71% (Micro Average: 96.92%)			
	True A	True B	Class Precision
Predicted Group A (Advanced Breast Cancer)	24	1	96.00%
Predicted Group B (NSAIDs)	1	39	97.50%
Class Recall	96.00%	97.50%	.

Table 4.5 Naive Bayes Performance metrics (NSAIDs VS Advanced Breast Cancer)

NB classifier predicted Misoprostol instead of Cisplatin to be a member of the ABCD Group (A). Cisplatin is the only member of Group (A) that is located in modularity number (4). However, all 40 members of Group (B) are exclusively localized in this particular modularity. Further graph analysis to modularity number (4) is presented in table (Table 4.7 Modularity (4) Graph analysis). Node degree and eigenvector centrality measures are listed and sorted in a decreasing order.

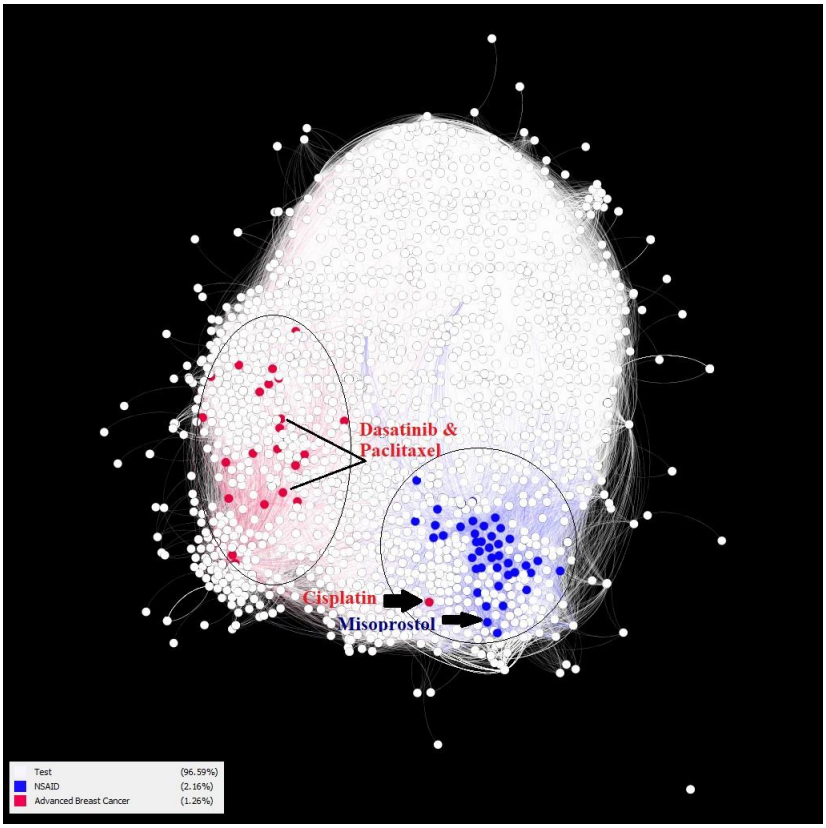

Figure 4.6 Advanced Breast Cancer Drugs VS NSAIDs

Modularity No.	ABCD Group (A)	Relative %	NSAIDs Group (B)	Relative %	Other Drugs	Grand Total
0	0	0.00%	0	0.00%	301	301
1	7	3.15%	0	0.00%	215	222
2	0	0.00%	0	0.00%	85	85
3	0	0.00%	0	0.00%	366	366
4	1	0.21%	40	8.35%	438	479
5	17	3.17%	0	0.00%	520	537
6	0	0.00%	0	0.00%	2	2

Table 4.6 Node Distribution Experiment (2)

Drug Name	Modularity No.	Node Degree	Eigenvector Centrality	Drug Group
Celecoxib	4	694	0.73	NSAID
Etoricoxib	4	530	0.57	NSAID
Diclofenac	4	514	0.54	NSAID
Rofecoxib	4	492	0.52	NSAID
Meloxicam	4	482	0.50	NSAID
Valdecoxib	4	478	0.50	NSAID
Lumiracoxib	4	455	0.45	NSAID
Indomethacin	4	460	0.45	NSAID
Ibuprofen	4	453	0.45	NSAID
Naproxen	4	449	0.45	NSAID
Mefenamic acid	4	439	0.43	NSAID
Piroxicam	4	439	0.43	NSAID
Phenylbutazone	4	433	0.43	NSAID
Oxaprozin	4	430	0.42	NSAID
Suprofen	4	430	0.42	NSAID
Tenoxicam	4	430	0.42	NSAID
Etodolac	4	430	0.42	NSAID
Ketoprofen	4	430	0.42	NSAID
Flurbiprofen	4	429	0.42	NSAID
Nimesulide	4	402	0.42	NSAID
Parecoxib	4	345	0.40	NSAID
Nabumetone	4	401	0.38	NSAID
Zomepirac	4	311	0.35	NSAID
Ketorolac	4	388	0.35	NSAID
Tiaprofenic acid	4	386	0.35	NSAID
Meclofenamic acid	4	384	0.35	NSAID
Tolmetin	4	384	0.35	NSAID
Fenoprofen	4	383	0.35	NSAID
Sulindac	4	389	0.35	NSAID
Lornoxicam	4	311	0.34	NSAID
Aceclofenac	4	310	0.34	NSAID
Oxyphenbutazone	4	359	0.32	NSAID
Niflumic Acid	4	272	0.28	NSAID
Fenbufen	4	255	0.25	NSAID
Ibuproxam	4	255	0.25	NSAID
Indoprofen	4	255	0.25	NSAID
Kebuzone	4	255	0.25	NSAID
Azapropazone	4	254	0.25	NSAID
Benoxaprofen	4	254	0.25	NSAID
Cisplatin	4	244	0.22	Advanced Breast Cancer
Misoprostol	4	101	0.10	NSAID

Table 4.7 Modularity (4) Graph analysis

Cisplatin shows node degree and eigenvector centrality values much closer to the members of the NSAIDs group. In addition, Misoprostol records significantly extreme low values compared to other members of the NSAIDs group.

Accuracy: 94.29% +/- 7.00% (Micro Average: 93.85%)			
	True A	True B	Class Precision
Predicted Group A (Advanced Breast Cancer)	22	1	95.65%
Predicted Group B (NSAIDs)	3	39	92.86%
Class Recall	88.00%	97.50%	

Table 4.8 Decision Tree Performance metrics (NSAIDs VS Advanced Breast Cancer)

Table (Table 4.8 Decision Tree Performance metrics (NSAIDs VS Advanced Breast Cancer)) reports the DT classifier results. Both NB and DT predicted Misoprostol as Group (A) member. Unlike NB, DT predicted two more drugs (Dasatinib and Paclitaxel) in addition to Cisplatin as Group (B) members.

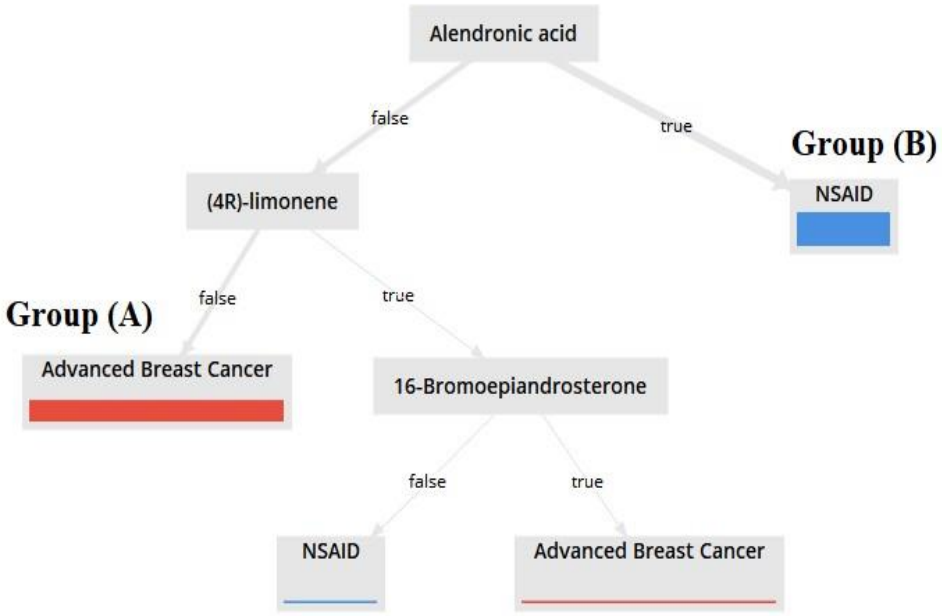

Figure 4.7 Decision Tree Result Experiment (2)

Figure (Figure 4.7 Decision Tree Result Experiment (2)) displays the Decision Tree results of the experiment (2). In this figure, the root node selected by the DT based on the maximum homogeneity in each leaf generated is located at the top (Alendronic acid). All nodes reported an interaction with Alendronic acid is classified as NSAIDs Group (B). The left arm of the tree further classifies the drugs based on another attribute (4R)-limonene. True or false indicates whether or not a drug has an interaction reported to the selected attribute. The line thickness infers the relative number of examples in each leaf compared to the overall examples.

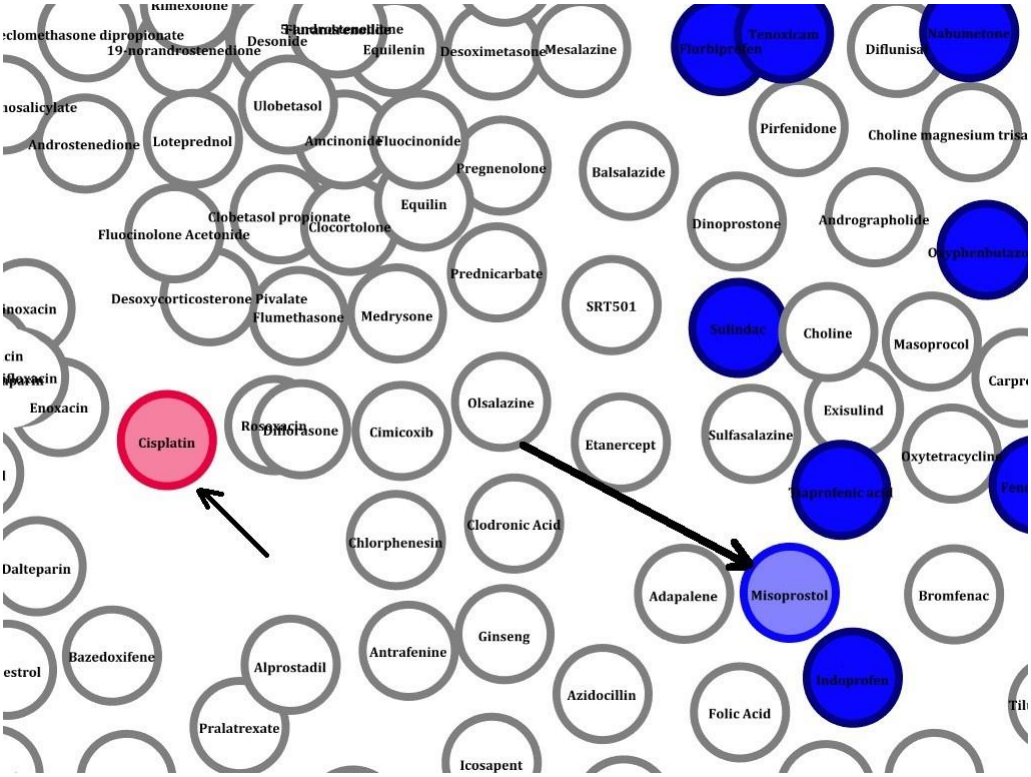

Figure 4.8 Misoprostol Position

Figure *(Figure 4.8 Misoprostol Position)* illustrates a close view to the position of Misoprostol and Cisplatin on the DDIs network graph compared to the high-level view in figure *(*

Figure 4.6 Advanced Breast Cancer Drugs VS NSAIDs).

No.	Drug Name	Label	Naive Bayes Prediction	Confidence	Supporting Evidence
1	Misoprostol	NSAID	Advanced Breast Cancer	1	(Lawson et al., 1994)
2	Cisplatin	Advanced Breast Cancer	NSAID	1	

Table 4.9 NB Predictions of Misoprostol Anti-Cancer Properties.

No.	Drug Name	Label	Decision Tree Prediction	Confidence	Supporting Evidence
1	Misoprostol	NSAID	Advanced Breast Cancer	1	(Lawson et al., 1994)
2	Cisplatin	Advanced Breast Cancer	NSAID	1	
3	Dasatinib	Advanced Breast Cancer	NSAID	1	
4	Paclitaxel	Advanced Breast Cancer	NSAID	1	

Table 4.10 Table 4.9 DT Predictions of Misoprostol Anti-Cancer Properties.

A summary in (Table 4.9 and Table 4.10) represents the classification results of both NB and DT between NSAIDs and Advanced Breast Cancer Group. (Lawson et al., 1994) mentioned a clinical evidence from the literature that confirms the anti-cancer properties of Misoprostol.

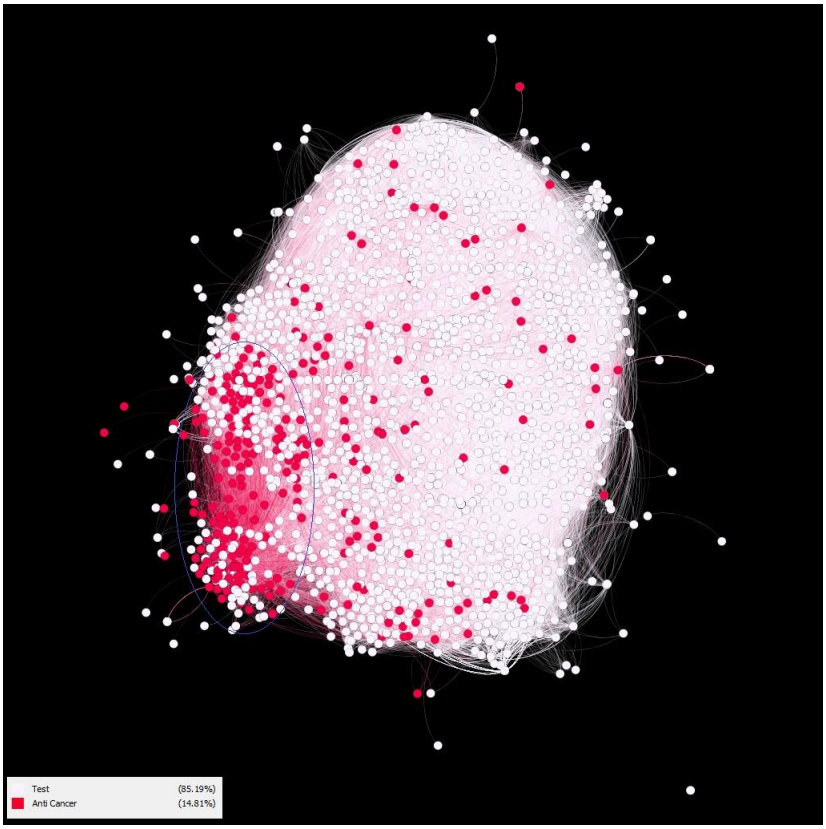

Figure 4.9 Anti-Cancer Drugs

4.1.3 Case Study: Anti-Cancer Drug Prediction.

This study tests the classification model on 2 drug groups. The first group includes 250 drugs approved for cancer treatment with clinically confirmed anti-cancer properties. All drug names and drug-drug interactions data are extracted from DrugBank (version 5.1.1) database. Figure (Figure 4.9 Anti-Cancer Drugs) displays the clustering of anti-cancer drugs in the DDIs network.

Anti-Cancer drug group is spread over the entire DDIs network. However, a significant number of this group is localized in a particular location in the graph. Table (Table 4.11 Anti-Cancer Drugs Node Distribution Case Study) displays the number of total nodes in each modularity and the component percentage within each modularity from each group. Modularity (1) contains the highest number of Anti-cancer drug group.

Modularity No.	Anti-cancer	Relative %	Other Drugs	Relative %	Grand Total
0	18	5.98%	283	94.02%	301
1	116	52.25%	106	47.75%	222
2	0	0.00%	85	100%	85
3	18	4.92%	348	95.08%	366
4	37	7.72%	442	92.28%	479
5	106	19.74%	431	80.26%	537
6	0	0.00%	2	100%	2

Table 4.11 Anti-Cancer Drugs Node Distribution Case Study

A total number of drugs within modularity (1) is 222 drugs, 52.25% of them belong to the Anti-Cancer group. Modularity (5) contains 106 Anti-Cancer drugs out of total 537 drugs. No Anti-Cancer drugs are present in modularity number (6, 2).

Accuracy: 92.02% +/- 27.10% (Micro Average: 92.02%)			
	True Test Class	True Anti-Cancer	Class Precision
Predicted. Test Class	1665	82	95.31%
Predicted. Anti-Cancer	76	168	68.57%
Class Recall	95.58%	67.20%	

Table 4.12 Decision Tree Performance Measures (Anti-Cancer VS Others)

Accuracy: 92.92% +/- 25.65% (Micro Average: 92.92%)			
	True Test Class	True Anti-Cancer	Class Precision
Predicted. Test Class	1662	62	96.41%
Predicted. Anti-Cancer	79	188	70.41%
Class Recall	95.46%	75.20%	

Table 4.13 Deep Learning Performance Measures (Anti-Cancer VS Others)

The confusion matrix of DT is presented in (Table 4.12 Decision Tree Performance Measures (Anti-Cancer VS Others)). DL shows a minor improvement of the precision and recall values of the Anti-Cancer group compared to DT's.

(Table 4.14, Table 4.15) provide a list of drugs with predicted anti-cancer properties. Each drug prediction is supported by clinical evidence from the literature that affirms the computational predictions. The reported clinical evidence along the prediction results could provide a guidance towards bringing more attention to those drugs, initiate controlled clinical trials to affirm its efficacy against cancer cell development.

No.	Drug Name	Prediction	Supporting Evidence
1	Clozapine	Anti-Cancer	(Fond, G and Macgregor, A and Attal, J and Larue, A and Brittner, M and Ducasse, D and Capdevielle, D, 2012)
2	Azathioprine	Anti-Cancer	(Rossi et al., 2018)
3	Inotuzumab ozogamicin	Anti-Cancer	(Dan et al., 2018)
4	Podofilox	Anti-Cancer	(Hu et al., 2018)
5	Albendazole	Anti-Cancer	(Cheong et al., 2018)
6	Cyproterone acetate	Anti-Cancer	(SIA, 2018)
7	Puromycin	Anti-Cancer	(Ueki & Hayman, 2018)
8	Sparsomycin	Anti-Cancer	(Moschetta et al., 2018)(Huang et al., 2018)
9	Interferon alfa-n1	Anti-Cancer	(Moschetta et al., 2018)
10	Lithium	Anti-Cancer	(Ozerdem et al., 2018) (Luo et al., 2018)
11	Benzphetamine	Anti-Cancer	(Singh & Sharma, 2018)
12	Anakinra	Anti-Cancer	(Wu et al., 2018) (Castaigne et al., 2018)(Tulotta & Ottewell, 2018)
13	Luliconazole	Anti-Cancer	(Ahmad et al., 2018)
14	Alosetron	Anti-Cancer	(Yin et al., 2018)
15	Resveratrol	Anti-Cancer	(Rauf et al., 2018) (Huminieck & Horbańczuk, 2018)
16	Cabergoline	Anti-Cancer	(Huang et al., 2018)
17	Colchicine	Anti-Cancer	(Rossi et al., 2018) (Ueki & Hayman, 2018)
18	Prasugrel	Anti-Cancer	(Bruno et al., 2018)
19	Hexestrol	Anti-Cancer	(Iwase et al., 2018)
20	Mycophenolic acid	Anti-Cancer	(Shah & Kharkar, 2018) (Fernández-Ramos et al., 2017)
21	Exisulind	Anti-Cancer	(Iwase et al., 2018) (Shi & Zhijian, 2018b) (Horinaka et al., 2014)
22	Metamizole	Anti-Cancer	(Malsy et al., 2017)
23	Pirfenidone	Anti-Cancer	(Li et al., 2018)
24	Ranibizumab	Anti-Cancer	(Castaigne et al., 2018) (Rossi et al., 2018)
25	Sulindac	Anti-Cancer	(Iwase et al., 2018) (Shi & Zhijian, 2018b) (Horinaka et al., 2014)

Table 4.14 Drugs with predicted Anti-Cancer Properties (a)

No.	Drug Name	Prediction	Supporting Evidence
26	Filgrastim	Anti-Cancer	(Norenberg, 2017) (Castaigne et al., 2018)
27	Daidzin	Anti-Cancer	(Graziani et al., 2018) (Huang et al., 2016)
28	Celecoxib	Anti-Cancer	(Hao et al., 2017)
29	Capsaicin	Anti-Cancer	(Clark & Lee, 2016) (Kang et al., 2016) (Granato et al.,
30	Misoprostol	Anti-Cancer	(Lawson et al., 1994)
31	Ethyl carbamate	Anti-Cancer	(Soni & Soman, 2018)
32	Suramin	Anti-Cancer	(Su et al., 2018)
33	Tetracycline	Anti-Cancer	(Lokeshwar et al., 2001)
34	Geldanamycin	Anti-Cancer	(Mayor-López et al., 2014)
35	Histamine	Anti-Cancer	(Garcia-Quiroz & Camacho, 2011)
35	Gemfibrozil	Anti-Cancer	(Lian et al., 2018)
36	Doxycycline	Anti-Cancer	(Lokeshwar et al., 2001)
37	Sirolimus	Anti-Cancer	(Jung et al., 2017)
38	Mebendazole	Anti-Cancer	(Rubin et al., 2018)

Table 4.15 Drugs with predicted Anti-Cancer Properties (b)

5 Conclusions and Future Work

Drug-Drug Interactions (DDIs) data could be utilized to build reliable drug profiles. In which, each drug profile is represented as a vector of features in a binary matrix of values [0, 1]. The binary matrix could be further analyzed by a Machine Learning classification model for drug repurposing to predict new drug indications.

This thesis addresses the significance of Machine Learning techniques in drug repositioning. The study investigates the role of DDIs information network as predictor features for novel drug properties. The study confirms that DDIs network clusters visualization provides significant information about drug features. In addition, network analysis supports the interpretation of unexplained drug behavior and suggest clues for drug repositioning. Furthermore, ML models successfully predicted drug properties based on DDIs information.

A case study about the prediction of anti-cancer drug properties successfully identified 76 drugs as potential drug repositioning candidates for cancer treatment. In this study, we selected the cutoff point of prediction significance as ≥ 95%. The candidates with confidence below the cutoff threshold were not mentioned. Extensive clinical literature survey supports the predicted features of selected drug candidates. In conclusion, applications of ML concepts and techniques provide the necessary tools to advance novel drugs discovery process.

Suggested ideas for future work continuation and improvement should be targeting the negative values in databases in order to distinguish the true absence of interactions from missing (not reported) or hidden (not yet discovered) cases. The negative values are considered a huge challenge for supervised machine learning techniques. We recommend investigating the role of applying association rules algorithm as a matrix completion technique for missing data, and evaluate its role on prediction algorithm performance measures.

In addition, the data type in DDIs database should be continuous or at least ordinal instead of binary [0, 1]. The binary data type is not accurately describing the interaction severity in the biological system.

6 References

Azuaje, F. (2012). Drug interaction networks: an introduction to translational and clinical applications. *Cardiovascular research*. The Oxford University Press, p. cvs289.

Bacciu, D., Lisboa, P. J. G., Mart\'\in, J. D., Stoean, R. & Vellido, A. (2018). Bioinformatics and Medicine in the Era of Deep Learning. *arXiv preprint arXiv:1802.09791*.

Barron, T. I., Connolly, R. M., Sharp, L., Bennett, K. & Visvanathan, K. (2011). Beta blockers and breast cancer mortality: a population-based study. *J Clin Oncol*, vol. 29(19), pp. 2635–2644.

Bastian, M., Heymann, S., Jacomy, M. & others. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*, vol. 8, pp. 361–362.

Berger, S. I. & Iyengar, R. (2009). Network analyses in systems pharmacology. *Bioinformatics*. Oxford University Press ({OUP}), vol. 25(19), pp. 2466–2472.

Brown, A. S., Kong, S. W., Kohane, I. S. & Patel, C. J. (2016). {ksRepo}: A Generalized Platform for Computational Drug Repositioning. *{BMC} Bioinformatics*. Springer Nature, vol. 17(1).

Bruno, A., Dovizio, M., Tacconelli, S., Contursi, A., Ballerini, P. & Patrignani, P. (2018). Antithrombotic agents and cancer. *Cancers*. Multidisciplinary Digital Publishing Institute, vol. 10(8), p. 253.

Caniza, H., Galeano, D. & Paccanaro, A. (2017). Mining the Biomedical Literature to Predict Shared Drug Targets in Drugbank. *Computer Conference (CLEI), 2017 XLIII Latin American*, pp. 1–5.

Cardelli, J. A., Circu, M. L., Dykes, S. S. & El-osta, H. E. (2018). Cancer Treatment Via Repositioned Tricyclic Anti-depressant-like Drugs As Anti-cancer Agents and New Combinations of Such Drugs. Google Patents.

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J. & Zhang, Y. (2015). Drug--target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics*. Oxford Univ Press, p. bbv066.

Cohen, M. E., Dembling, B. & Schorling, J. B. (2002). The association between schizophrenia and cancer: a population-based mortality study. *Schizophrenia Research*. Elsevier, vol. 57(2–3), pp. 139–146.

Coussens, L. M. & Werb, Z. (2002). Inflammation and cancer. *Nature*. Springer Nature, vol. 420(6917), pp. 860–867.

Fond, G and Macgregor, A and Attal, J and Larue, A and Brittner, M and Ducasse, D and Capdevielle, D. (2012). Antipsychotic drugs: pro-cancer or anti-cancer? A systematic review. *Medical hypotheses*. Elsevier, vol. 79(1), pp. 38–42.

Franciosi, M., Lucisano, G., Lapice, E., Strippoli, G. F. M., Pellegrini, F. & Nicolucci, A. (2013). Metformin Therapy and Risk of Cancer in Patients with Type 2 Diabetes: Systematic Review. *PLOS ONE*. Public Library of Science, vol. 8(8).

Garcia-Quiroz, J. & Camacho, J. (2011). Astemizole: an old anti-histamine as a new promising anti-cancer drug. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*. Bentham Science Publishers, vol. 11(3), pp. 307–314.

Gil-Ad, I., Shtaif, B., Levkovitz, Y., Dayag, M., Zeldich, E. & Weizman, A. (2004). Characterization of phenothiazine-induced apoptosis in neuroblastoma and glioma cell lines. *Journal of Molecular Neuroscience*. Springer, vol. 22(3), pp. 189–198.

Hanusova, V., Skalova, L., Kralova, V. & Matouskova, P. (2015). Potential anti-cancer drugs commonly used for other indications. *Current cancer drug targets*. Bentham Science Publishers, vol. 15(1), pp. 35–52.

Hao, M., Bryant, S. H. & Wang, Y. (2017). Predicting Drug-target Interactions by Dual-network Integrated Logistic Matrix Factorization. *Scientific Reports*. Springer Nature, vol. 7, p. 40376.

Hattori, M., Okuno, Y., Goto, S. & Kanehisa, M. (2003). Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. *Journal of the American Chemical Society*. ACS Publications, vol. 125(39), pp. 11853–11865.

Hosono, K., Endo, H., Takahashi, H., Sugiyama, M., Sakai, E., Uchiyama, T., Suzuki, K., Iida, H., Sakamoto, Y., Yoneda, K., Koide, T., Tokoro, C., Abe, Y., Inamori, M., Nakagama, H. & Nakajima, A. (2010). Metformin Suppresses Colorectal Aberrant Crypt Foci in a Short-term Clinical Trial. *Cancer Prevention Research*. American Association for Cancer Research, vol. 3(9), pp. 1077–1083.

Hu, T.-M. & Hayton, W. L. (2011). Architecture of the drug--drug interaction network. *Journal of clinical pharmacy and therapeutics*. Wiley Online Library, vol. 36(2), pp. 135–143.

Huang, J., Zhao, D., Liu, Z. & Liu, F. (2018). Repurposing psychiatric drugs as anti-cancer agents. *Cancer Letters*. Elsevier {BV}, vol. 419, pp. 257–265.

Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P. & Agarwal, P. (2013). Computational Drug Repositioning: From Data to Therapeutics. *Clinical Pharmacology & Therapeutics*. Springer Nature, vol. 93(4), pp. 335–341.

Jamal, S., Goyal, S., Shanker, A. & Grover, A. (2017). Predicting Neurological Adverse Drug Reactions Based on Biological, Chemical and Phenotypic Properties of Drugs Using Machine Learning Models. *Scientific Reports*. Nature Publishing Group, vol. 7(1), p. 872.

Jin, B., Yang, H., Xiao, C., Zhang, P., Wei, X. & Wang, F. (2017). Multitask Dyadic

Prediction and Its Application in Prediction of Adverse Drug-drug Interaction. *AAAI*, pp. 1367–1373.

Jung, K. S., Lee, J., Park, S. H., Park, J. O., Park, Y. S., Lim, H. Y., Kang, W. K. & Kim, S. T. (2017). Pilot study of sirolimus in patients with {PIK}3CA mutant/amplified refractory solid cancer. *Molecular and Clinical Oncology*. Spandidos Publications, vol. 7(1), pp. 27–31.

Keane, H., Ryan, B. J., Jackson, B., Whitmore, A. & Wade-Martins, R. (2015). Protein-protein interaction networks identify targets which rescue the MPP+ cellular model of Parkinson's disease. *Scientific reports*. Nature Publishing Group, vol. 5, p. 17004.

Krzywinski, M., Birol, I., Jones, S. J. M. & Marra, M. A. (2011). Hive plots—rational approach to visualizing networks. *Briefings in bioinformatics*. Oxford University Press, vol. 13(5), pp. 627–644.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V. & others. (2013). Drugbank 4.0: Shedding New Light on Drug Metabolism. *Nucleic acids research*. Oxford University Press, vol. 42(D1), pp. D1091--D1097.

Lee, M. S., Johansen, L., Zhang, Y., Wilson, A., Keegan, M., Avery, W., Elliott, P., Borisy, A. A. & Keith, C. T. (2007). The novel combination of chlorpromazine and pentamidine exerts synergistic antiproliferative effects through dual mitotic action. *Cancer research*. AACR, vol. 67(23), pp. 11359–11367.

Lian, X., Wang, G., Zhou, H., Zheng, Z., Fu, Y. & Cai, L. (2018). Anticancer Properties of Fenofibrate: A Repurposing Use. *Journal of Cancer*. Ivyspring International Publisher, vol. 9(9), pp. 1527–1537.

Lien, E. J. and Ou, Xing-chang. (1985). Carcinogenicity of some anticancer drugs—A

survey. *Journal of Clinical Pharmacy and Therapeutics*. Wiley Online Library, vol. 10(3), pp. 223–242.

Liu, L., Chen, L., Zhang, Y.-H., Wei, L., Cheng, S., Kong, X., Zheng, M., Huang, T. & Cai, Y.-D. (2016). Analysis and Prediction of Drug{\textendash}drug Interaction by Minimum Redundancy Maximum Relevance and Incremental Feature Selection. *Journal of Biomolecular Structure and Dynamics*. Informa {UK} Limited, vol. 35(2), pp. 312–329.

Liu, Z., Fang, H., Reagan, K., Xu, X., Mendrick, D. L., Slikker, W. & Tong, W. (2013). In silico drug repositioning--what we need to know. *Drug discovery today*. Elsevier, vol. 18(3), pp. 110–115.

Lokeshwar, B., Escatel, E. & Zhu, B. (2001). Cytotoxic Activity and Inhibition of Tumor Cell Invasion by Derivatives of a Chemically Modified Tetracycline {CMT}-3 ({COL}-3). *Current Medicinal Chemistry*. Bentham Science Publishers Ltd., vol. 8(3), pp. 271–279.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L. & Zeng, J. (2017). A Network Integration Approach for Drug-target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. *Nature Communications*. Springer Nature, vol. 8(1).

Mayor-López, L., Tristante, E., Carballo-Santana, M., Carrasco-Garc\'\ia, E., Grasso, S., Garc\'\ia-Morales, P., Saceda, M., Luján, J., Garc\'\ia-Solano, J., Carballo, F., de Torre, C. & Mart\'\inez-Lacaci, I. (2014). Comparative Study of 17-{AAG} and {NVP}-{AUY}922 in Pancreatic and Colorectal Cancer Cells: Are There Common Determinants of Sensitivity? *Translational Oncology*. Elsevier {BV}, vol. 7(5), pp. 590–604.

Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L. & Zheng, J. (2012). Drug--target interaction prediction by learning from local information and neighbors. *Bioinformatics*. Oxford University Press, vol. 29(2), pp. 238–245.

Mohammadian, M., Zeynali, S., Azarbaijani, A. F., Ansari, M. H. K. & Kheradmand, F. (2017). Cytotoxic effects of the newly-developed chemotherapeutic agents 17-AAG in combination with oxaliplatin and capecitabine in colorectal cancer cell lines. *Research in pharmaceutical sciences*. Wolters Kluwer--Medknow Publications, vol. 12(6), p. 517.

Munir, A., Elahi, S. & Masood, N. (2018). Clustering Based Drug-drug Interaction Networks for Possible Repositioning of Drugs against {EGFR} Mutations: Clustering Based {DDI} Networks for {EGFR} Mutations. *Computational Biology and Chemistry*. Elsevier {BV}, vol. 75, pp. 24–31.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*. SIAM, vol. 45(2), pp. 167–256.

Nordenberg, J., Fenig, E., Landau, M., Weizman, R. & Weizman, A. (1999). Effects of psychotropic drugs on cell proliferation and differentiation. *Biochemical pharmacology*. Elsevier, vol. 58(8), pp. 1229–1236.

Oliveras-Ferraros, C., Cufí, S., Vazquez-Martin, A., Torres-Garcia, V. Z., Barco, S. Del, Martin-Castillo, B. & Menendez, J. A. (2011). Micro(mi)RNA expression profile of breast cancer epithelial cells treated with the anti-diabetic drug metformin: Induction of the tumor suppressor miRNA let-7a and suppression of the TGFβ-induced oncomiR miRNA-181a. *Cell Cycle*. Taylor & Francis, vol. 10(7), pp. 1144–1151.

Peer, D. & Margalit, R. (2006). Fluoxetine and reversal of multidrug resistance. *Cancer letters*. Elsevier, vol. 237(2), pp. 180–187.

Peng, L., Liao, B., Zhu, W. & Li, K. (2015). Predicting Drug-Target Interactions with Multi-information Fusion. *IEEE journal of biomedical and health informatics*. IEEE, vol. 21(2), pp. 561–572.

Peng, L., Zhu, W., Liao, B., Duan, Y., Chen, M., Chen, Y. & Yang, J. (2017). Screening

Drug-target Interactions with Positive-unlabeled Learning. *Scientific Reports*. Nature Publishing Group, vol. 7(1), p. 8087.

Qi, L. & Ding, Y. (2013). Potential antitumor mechanisms of phenothiazine drugs. *Science China Life Sciences*. Springer, vol. 56(11), pp. 1020–1027.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*. Springer, vol. 1(1), pp. 81–106.

Raja, K., Patrick, M., Elder, J. T. & Tsoi, L. C. (2017). Machine Learning Workflow to Enhance Predictions of Adverse Drug Reactions ({ADRs}) through Drug-gene Interactions: Application to Drugs for Cutaneous Diseases. *Scientific Reports*. Springer Nature, vol. 7(1).

Rubin, J., Mansoori, S., Blom, K., Berglund, M., Lenhammar, L., Andersson, C., Loskog, A., Fryknäs, M., Nygren, P. & Larsson, R. (2018). Mebendazole stimulates CD14+ myeloid cells to enhance T-cell activation and tumour cell killing. *Oncotarget*. Impact Journals, LLC, vol. 9(56), p. 30805.

Sawada, R., Iwata, M., Tabei, Y., Yamato, H. & Yamanishi, Y. (2018). Predicting Inhibitory and Activatory Drug Targets by Chemically and Genetically Perturbed Transcriptome Signatures. *Scientific reports*. Nature Publishing Group, vol. 8(1), p. 156.

Setoain, J., Franch, M., Mart\'\inez, M., Tabas-Madrid, D., Sorzano, C. O. S., Bakker, A., Gonzalez-Couto, E., Elvira, J. & Pascual-Montano, A. (2015). {NFFinder}: An Online Bioinformatics Tool for Searching Similar Transcriptomics Experiments in the Context of Drug Repositioning. *Nucleic Acids Research*. Oxford University Press ({OUP}), vol. 43(W1), pp. W193--W199.

Shi, Y. & Zhijian, S. (2018a). Applications for Nicardipine in Preparing Anti-lung Cancer Products. Google Patents.

Shi, Y. & Zhijian, S. (2018b). Applications For Sulindac In Preparing Anti-lung Cancer

Products. Google Patents.

Shi, Y. & Zhijian, S. (2018c). Applications Of Desogestrel In The Preparation Of Anti-colon Cancer/breast Cancer Er-negative Ah Receptor-positive Products. Google Patents.

Singh, N. & Sharma, B. (2018). Toxicological Effects of Berberine and Sanguinarine. *Frontiers in molecular biosciences*. Frontiers, vol. 5, p. 21.

Soni, R. & Soman, S. S. (2018). Design and synthesis of aminocoumarin derivatives as {DPP}-{IV} inhibitors and anticancer agents. *Bioorganic Chemistry*. Elsevier {BV}, vol. 79, pp. 277–284.

Su, L., Vasile, S., Smith, L. & Leng, F. (2018). Identification of Suramin as a Potent and Specific Inhibitor of the Mammalian High Mobility Group Protein at-Hook 2 (HMGA2)-DNA Interactions. *Biophysical Journal*. Elsevier, vol. 114(3), p. 443a.

Taylor, M. A., Sossey-Alaoui, K., Thompson, C. L., Danielpour, D. & Schiemann, W. P. (2013). TGF-β upregulates miR-181a expression to promote breast cancer metastasis. *The Journal of Clinical Investigation*. The American Society for Clinical Investigation, vol. 123(1), pp. 150–163.

Udrescu, L., Sbârcea, L., Top\^\irceanu, A., Iovanovici, A., Kurunczi, L., Bogdan, P. & Udrescu, M. (2016). Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing. *Scientific Reports*. Nature Publishing Group, vol. 6(1).

Wang, Y. & Zeng, J. (2013). Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*. Oxford Univ Press, vol. 29(13), pp. i126--i134.

Weiss, S. M. & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc.

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. (2008). Prediction of drug--target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. Oxford University Press, vol. 24(13), pp. i232--i240.

Yde, C. W., Clausen, M. P., Bennetzen, M. V, Lykkesfeldt, A. E., Mouritsen, O. G. & Guerra, B. (2009). The antipsychotic drug chlorpromazine enhances the cytotoxic effect of tamoxifen in tamoxifen-sensitive and tamoxifen-resistant human breast cancer cells. *Anti-cancer drugs*. LWW, vol. 20(8), pp. 723–735.

Yin, W., Gao, C., Xu, Y., Li, B., Ruderfer, D. M. & Chen, Y. (2018). Learning Opportunities for Drug Repositioning via GWAS and PheWAS Findings. *AMIA Summits on Translational Science Proceedings*. American Medical Informatics Association, vol. 2017, p. 237.

Yoo, S., Noh, K., Shin, M., Park, J., Lee, K.-H., Nam, H. & Lee, D. (2018). In Silico Profiling of Systemic Effects of Drugs to Predict Unexpected Interactions. *Scientific Reports*. Springer Nature, vol. 8(1).

Yu, L., Ma, X., Zhang, L., Zhang, J. & Gao, L. (2016). Prediction of New Drug Indications Based on Clinical Data and Network Modularity. *Scientific Reports*. Springer Nature, vol. 6(1).

Zhang, C. & Chu, M. (2018). Leflunomide: A promising drug with good antitumor potential. *Biochemical and Biophysical Research Communications*. Elsevier {BV}, vol. 496(2), pp. 726–730.

Zhang, J. & Huan, J. (2010). Analysis of Network Topological Features for Identifying Potential Drug Targets. *Proc 9th Intl Workshop Data Mining Bioinformatics (BioKDD'10), Washington DC, July 2010*.

Zhang, K., Chen, X., Wang, H. & Wang, Y. (2018). 'Application of clinical diagnosis and treatment data of coronary heart disease based on association rules'. , in *Recent*

Developments in Data Science and Business Analytics. Springer International Publishing, pp. 367–372.

Zhao, K. & So, H.-C. (2018). Drug Repositioning for Schizophrenia and Depression/anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *{IEEE} Journal of Biomedical and Health Informatics*. Institute of Electrical and Electronics Engineers ({IEEE}), p. 1.