Detecting Adverse Drug Events Using a Deep neural network Model

Saminur Islam
West Virginia University, si0019@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
Islam, Saminur, "Detecting Adverse Drug Events Using a Deep neural network Model" (2022). Graduate Theses, Dissertations, and Problem Reports. 11159.
https://researchrepository.wvu.edu/etd/11159

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
Detecting Adverse Drug Events using a Deep Neural Network Model

By

Saminur Islam

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science and Electrical Engineering
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the West Virginia University

Dr. Donald Adjeroh
Dr. Gianfranco Doretto
Dr. Wanhong Zheng
Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2022

Keywords:[Protein Sequence, Protein Structure, PS3N, PSSCSSN, Chemical Structure, Neural network Model]

©2022 Saminur Islam
Detecting Adverse Drug Events using a Deep Neural Network Model

Saminur Islam

Adverse drug events represent a key challenge in public health, especially with respect to drug safety profiling and drug surveillance. Drug-drug interactions represent one of the most popular types of adverse drug events. Most computational approaches to this problem have used different types of drug-related information utilizing different types of machine learning algorithms to predict potential interactions between drugs. In this work, our focus is on the use of genetic information about the drugs, in particular, the protein sequence and protein structure of drug protein targets to predict potential interactions between drugs. We collected information on drug-drug interactions (DDIs) from the DrugBank database and divided them into multiple datasets based on the type of information, such as, chemical structure, protein targets, side effects, pathways, protein-protein interactions, protein structure, information about indications. We proposed a similarity-based Neural Network framework called protein sequence-structure similarity network (S3N), and used this to predict the novel DDI’s. The drug-drug similarities are computed using different categories of drug information based on multiple similarity metrics. We compare the results with those from the state-of-the-art methods on this problem. Our results show that proposed method is quite competitive, at times outperforming the state-of-the-art. Our performance evaluations on different datasets showed the predictive performance as follows: Precision 91%-98%, Recall 90%-96%, F1 Score 86%-95%, AUC 88%-99% Accuracy 86%-95%. To further investigate the reliability of the proposed method, we utilize 158 drugs related to cardiovascular disease to evaluate the performance of our model
and find out the new interactions among the drugs. Our model showed 90% accuracy of detecting the existing drug interactions and identified 60 new DDI’s for the cardiovascular drugs. Our evaluation demonstrates the effectiveness of S3N in predicting DDI’s.
DEDICATION

I would like to dedicate this thesis to my mom for her incredible love and support. Because I believe that she is the real backbone of our family, this is to appreciate her selfless hard work and efforts towards the family.

Furthermore, I dedicate it to my dad to raise me like a son and give me wings to fly. To my Wife, for always trusting me and supporting me in my hard times, without her encouragement, nothing would have been easy. And to my entire family for their unconditional affection towards me.
AKNOWLEDGEMENTS

I would like to sincerely express my most profound gratitude towards my supervisor Dr. Donald Adjeroh, whose input helped me immensely. With his input, I was able to look at my research with a different perspective and a more critical eye.

Secondly, I would like to express my gratitude to my thesis committee members for their beneficial advice and suggestions for my thesis.

I would also like to thank my friends and Colleagues for always encouraging and supporting me.

I humbly extend my thanks to the Department of Computer Science and Electrical Engineering and all concerned people who helped me in this regard.
# TABLE OF CONTENTS

**ABSTRACT**  
ii

**DEDICATION**  
iv

**AKNOWLEDGEMENTS**  
v

**LIST OF TABLES**  
viii

**LIST OF FIGURES**  
ix

**LIST OF ABBREVIATIONS**  
xi

## 1 Introduction
1.1 Significance of DDI problem  
1
1.2 Steps taken by the researchers  
2
1.3 Problem Statement  
2
1.4 Objective  
4
1.5 Thesis Organization  
4

## 2 Related Works  
6

## 3 Methodology  
9
3.1 Important Terminologies  
9
3.1.1 Drug Protein targets  
10
3.1.2 Drug Action  
10
3.1.3 Drug Pairwise Similarity vs Drug Action Similarity  
10
3.1.4 Drug Similarity space  
11
3.1.5 K-mer  
11
3.1.6 Suffix Array  
11
3.2 Distance Matrices  
12
3.2.1 Cosine Similarity (CS)  
12
3.2.2 Levenshtein Distance (L)  
12
3.2.3 Jensen Shannon (JS) divergence(JSD)  
13
3.2.4 Euclidean Distance  
13
3.2.5 Tanimoto Coefficient  
14
3.3 Similarity Matrices  
14
3.3.1 Protein Sequence and Structure Similarity Matrices  
14
  3.3.1.1 Protein Sequence Similarity Matrices  
15
  3.3.1.2 Protein Structure Similarity Matrices  
15
3.4 Protein Sequence-Structure Similarity Network (PS3N)  
17
3.5 Neural Network Model  
18
# Application to Cardiovascular Drugs

4.1 Similarity Metrics based on other drug Information .................................................. 21
  4.1.1 Protein-Protein Interaction (PPI) Network ......................................................... 21
  4.1.2 Pathway Network ............................................................................................... 22
    4.1.2.1 Weighted Similarity of Drug pathway Network ................................................. 23
  4.1.3 Protein-Protein Alignment Similarity Network .................................................... 24
  4.1.4 Side-Effects Similarity Network ........................................................................... 25
  4.1.5 Chemical Structure Drug Similarity Network ...................................................... 25

4.2 Protein Sequence, Structure and Chemical Structure Similarity Network (PSSCSSN) .................. 26

4.3 Neural Network Model .............................................................................................. 27

# Experiments and Results

5.1 Datasets ...................................................................................................................... 28

5.2 Performance Evaluation of PS3N Model .................................................................... 29
  5.2.1 Experimental setups ............................................................................................ 30
  5.2.2 Comparing performance with State-of-the-art Algorithms .................................. 30
  5.2.3 Predicted DDI Network Diagram Using PS3N .................................................... 31

5.3 Impact of algorithmic parameters .............................................................................. 33

5.4 Results on Cardiovascular Drugs .............................................................................. 36
  5.4.1 Predicted Interactions of cardiovascular drugs ..................................................... 37
    5.4.1.1 PS3N DDI Network for cardiovascular drugs ................................................. 45
    5.4.1.2 PSSCSSN DDI Network for cardiovascular drugs ........................................... 46

# Discussion

6. Discussion .................................................................................................................... 48

# Conclusion

7. Conclusion .................................................................................................................. 50

REFERENCES ................................................................................................................... 52
LIST OF TABLES

5.2.1 Performance of PS3N using similarity matrices based on protein sequences .......................................................... 31
5.2.2 Performance of PS3N using similarity matrices based on protein structure.............................................................. 32
5.2.3 Results using combined protein sequence and protein structure similarity matrices................................................. 32
5.2.4 Performance comparison of different methods on DS1 from [41]. We obtained information on 469 drugs for protein sequences, and on 414 drugs for protein structure. The first Six rows are from [41] to compare the results from our model ................................................. 33
5.2.5 Performance comparison of different methods on the DS2 Dataset from [41]. We obtained information on 585 drugs for protein sequences, and on 504 drugs for protein structure. The first Six rows are from [41] to compare the results from our model .............. 34
5.3.1 Results of PS3N with variation on algorithmic parameters .............. 35
5.4.1 Performance comparison of PS3N model and PSSCSSN model ........ 37
5.4.2 Sample of predicted new DDI’s for cardiovascular drugs using PSSCSSN Model .......................................................... 38
5.4.3 Discovered new Inter cardiovascular DDI’s using PSSCSSN Model .. 39
LIST OF FIGURES

3.4.1 Proposed Protein Sequence-Structure Similarity Network (PS3N) model for predicting adverse drug events. Using the method of Similarity Network Fusion (SNF) we create a single $N \times N$ fusion matrix for $N$ drugs. From the fusion matrix, we compute the feature vectors for each pair of drugs. In this way we will have possible $\binom{N}{2}$ rows, and each row will have $N$ columns as features. These feature vectors are then fed into a multi-layer perceptron model. For protein sequence similarity network, the number of hidden layers would reduce to 3 since we have less number of drugs. ................................. 18

5.2.1 Drug Drug Interaction Similarity Network Diagram. We have used a threshold value $\geq 0.7$ in PS3N Network to discard a large chunk of network diagram. If Any pair of drugs have predicted labels as 1, we consider an edge between the drugs. In these way we created a Network diagram for PS3N Model using PS3N Network. ............... 35

5.4.1 PS3N network diagram for the cardiovascular drugs. In this network, there are only 87 cardiovascular drugs are available among the 489 drugs we considered for the network diagram. This network is representing the overall interactions based on PS3N network. The red lines are indicating the new interactions. ................................. 45

5.4.2 This diagram is representing the network for drug Alteplase. The green node representing Alteplase and the red dotted lines are representing the new inter-cardiovascular drug interactions. Among the 87 cardiovascular drugs, there are new 6 interactions for Alteplase. .... 46
5.4.3 PSSCSSN Network diagram for a particular cardiovascular drug Pravastatin. The network contains an overall 722 drugs. It is very difficult to show every interactions from all the nodes. So we considered a subset of interactions which are newly found and interacting with Pravastatin. Green lines represents the inter-cardiovascular interactions with Pravastatin.

5.4.4 Cardiovascular interaction diagram for Pravastatin. All other drugs interactions are discarded for better visualization and understanding.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TPR</td>
<td>True Positive Rate</td>
</tr>
<tr>
<td>FPR</td>
<td>False Positive Rate</td>
</tr>
<tr>
<td>FNR</td>
<td>False Negative Rate</td>
</tr>
<tr>
<td>TNR</td>
<td>True Negative Rate</td>
</tr>
<tr>
<td>FRR</td>
<td>False Rejection Rate</td>
</tr>
<tr>
<td>FNMR</td>
<td>False Non Match Rate</td>
</tr>
<tr>
<td>FAR</td>
<td>False Rejection Rate</td>
</tr>
<tr>
<td>FMR</td>
<td>False Match Rate</td>
</tr>
<tr>
<td>EER</td>
<td>Equal Error Rate</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>KNN</td>
<td>K Nearest Neighbors</td>
</tr>
<tr>
<td>NB</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi Layer Perceptron</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Network</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

Given the increasing number of medications that are being consumed concurrently by individuals, it is becoming more and more important to know more about the drugs we take. With this increased potential for polypharmacy, there is a corresponding increase in the chance of adverse events involving medications. It has been revealed that drugs may interact with others when they are taken together, and unexpected drug-drug interactions (DDIs) may lead to unexpected adverse drug events[48, 18]. The sheer number of people taking more than one medication in a given day has made the issue of drug-drug interactions a major public health problem. So, the more DDIs we know, the more we could take necessary measures to avoid unforeseen adverse drug events.

1.1 Significance of DDI problem

Recent advances in biomedical research have generated a large volume of drug-related data. To effectively handle this enormous amount of data, many initiatives have been introduced to help researchers make sense of the massive data sets. Drug knowledge bases such as Drug Bank, SIDER, PDB, STITCH, SMILES, Protein Data Bank, and others have emerged as a result of this. These knowledge bases contain a variety of drug-related information, such as genetic sequences, protein structures, drug side effects, chemical structures, drug indications, etc. Thus, several approaches have been proffered to utilize the information from these different sources to predict potential interactions between drugs [29, 31, 57, 43, 50]. DrugBank is perhaps one of
the most credible databases of known DDIs \cite{17,20,33} and contains information on over 300,000 DDIs. However, the number of drug-drug interactions is less than 1% of the total possible drug pairs that exist in DrugBank. DDI’s are known as the unwanted side effects resulting from the concurrent consumption of two or more drugs\cite{36,47,35}. When a doctor prescribes several drugs simultaneously for a patient, this may cause irreparable side effects. The effects of drugs on each other may lead to other illnesses or even death. These side effects are particularly noticeable in the elderly, or in persons with challenging diseases, such as cancer patients, who take many different drugs daily. Given the relevance of this in an individual’s health, and to public health in general, there is a critical need for more accurate and effective computational methods for understanding DDIs and how to predict them.

1.2 Steps taken by the researchers

However, researchers have developed several computational prediction strategies to address adverse drug events by detecting DDIs in recent years. Among the different approaches, machine learning-based DDI prediction has been used popularly for predicting the DDI’s with reduced time and cost\cite{32,33,34,17,38,59,62,61}. Many machine learning-based DDI prediction methods have been proposed, and are roughly classified into four categories: similarity-based methods, network-based methods, matrix factorization-based methods, and ensemble learning-based methods\cite{12}. The similarity-based methods are one major category among these methods.

1.3 Problem Statement

A drug interaction occurs when two (or more) drugs interact, or when a drug interacts with food, beverage, or supplement. Drug interactions can reduce the effectiveness of your medication, induce unanticipated side effects, or boost the impact of a drug. Some drug interactions can be dangerous to your health. It’s possible that reading the label every time you use a nonprescription or prescription drug, as well as learning
about drug interactions, will save your life[1]. A drug-drug interaction may increase or decrease the effects of one or both drugs. Clinically significant interactions are often predictable and usually undesired (see Some Drugs With Potentially Serious Drug-Drug Interactions)[66]. Adverse effects or therapeutic failure may result. Rarely, clinicians can use predictable drug-drug interactions to produce a desired therapeutic effect[52]. Moreover, Drug interaction is a leading cause of adverse drug events and a major obstacle for current clinical practice[66]. Adverse drug events (ADEs), the unintended drug side effects, have led to a major public health burden. In the United States alone, > 500,000 serious ADEs were reported annually to the US Food and Drug Administration (FDA) during the past 5 years [2]. Here we review the recent developments in addressing the challenge of adverse drug events within the range of different domains. And also we construct a set of data sets to search for new DDI's.

Patients with cardiovascular diseases (CVD) are at high risk of experiencing DDIs[6]. A scientifically driven and reasonable approach to drug interactions can reduce the risk of harmful effects and enhance patient outcomes. Cardiovascular drugs are used to treat a variety of illnesses, but there are often significant variances in treatment response and dosage requirements between patients. Treatment that works for one person may be useless or even hazardous for another[46]. Patients with pre-existing cardiac problems are frequently prescribed multiple drugs, such as anticoagulants and antiarrhythmics, which can interact with a range of chemotherapies in real and prospective ways. The focus of treatment in cardio-oncology has evolved from reactive to proactive care as the profession has progressed[70]. Because the therapeutic window for most chemotherapies is so short, these drug-drug interactions that may improve or decrease chemotherapeutic efficacy or predispose patients to major unexpected side effects should be given considerable consideration[70]. So a system of predicting the interaction among drugs in cardiovascular disease and planning an individual patients’ drug therapies is in dire need.
1.4 Objective

After looking through the published work for predicting drug-drug interactions, we discovered that none of the algorithms meet the goals. Because detecting DDIs in a clinical test is extremely challenging. Many DDIs are dose-dependent, and the nature of the drug approval process, as well as inherent genetic and demographic diversity, can all cause DDI Recognition to be delayed[45]. The primary objective of most machine learning and deep learning approaches is to identify interactions. In the current research on DDI prediction, there is surprisingly little information on medications used to address the problem. The main objective of most machine learning and deep learning approaches is to identify interactions. The current research on DDI prediction has utilized only a small amount of drug information. Our goal is to create a deep neural network framework that takes advantage of all available pharmacological information, including chemical structure, protein structure, protein sequence, side effects, interactions between proteins, pathways, etc.

Following the development of a deep learning framework for predicting potential novel DDIs, we use the methodology for the cardiovascular disease problem. Considering how many drugs are available on the market today and how many people take different medications based on prescriptions, it is likely that there will be multiple interactions, including interactions with cardiovascular drugs and interactions with other medications. Our study will look at both types of interactions and see how effective the clinical trial is at improving patients’ drug regimes.

1.5 Thesis Organization

In this thesis, we investigate how this can be predicted from different categories of drug information. We organize our work as follows: In the next chapter, we review related work. In the third chapter, we discussed our proposed methodology. Chapter four describes the extension of our work on the DDI prediction model and its application in cardiovascular drug interaction prediction. In Chapter five, we describe
our experimental studies and observed results for drug-drug interaction prediction and how we plan to use our proposed method to discover drug interactions for cardiovascular drugs. Chapter six would represent the discussions and limitations of our proposed methodology and possible improvements. Finally, Chapter seven will present the Conclusions and Future Work.
CHAPTER 2

Related Works

Most existing approaches for DDI prediction are based on different properties of the drug compound, such as its chemical structure, side effects, drug-target relationship, and many more. DDIs can be identified with in vivo models using high-throughput screening [25]. However, the price of such procedures is relatively high, and testing large numbers of drug combinations is not practical [23]. To reduce the number of possible drug combinations, numerous computational approaches have been proposed [28, 56, 55, 10, 67, 30, 42, 44]. In some of these computational approaches, drug-target networks are constructed, and DDIs are detected by measuring the strength of network connections [67], or by identifying drug pairs that share drug targets or drug pathways, for instance, using the random walk algorithm [30].

Some computational approaches have used the structural similarity and side effect similarities of drug pairs. For example, Gottlieb et al. proposed the Inferring Drug Interactions (INDI) method, which predicts novel DDIs from chemical and side effect similarities of known DDIs [28]. Vilar et al. used similarities of fingerprints, target genes, and side effects of drug pairs [56, 55]. Cheng et al. constructed features from the Simplified Molecular-Input Line-Entry System (SMILES) data and side effect similarity of drug pairs and applied support vector machines to predict DDIs [10]. Zhang et al. constructed a network of drugs based on structural and side effect similarities and applied a label propagation algorithm to identify DDIs [67]. Recently, Ryu et al. proposed DeepDDI, a computational framework that calculates structural similarity profiles (SSP) of DDIs, reduces features using principal component analysis (PCA), and feeds them to a feed-forward deep neural network [49]. The platform
generated 86 labeled pharmacological DDI effects, so DeepDDI [26] is basically a multi-classification (multi-label classification) model.

There are some machine learning–based methods apply KNN [8], SVM [8], logistic regression [10, 29, 54], decision tree [10], naïve Bayes [10], and network-based label propagation [67] and random walk [69] or matrix factorization [65] to detect DDIs. These methods are based on drug properties, such as chemical structure [10, 29, 67, 69, 8], targets [10, 29, 54], Anatomical Therapeutic Chemical classification (ATC) codes [10, 29, 8], side effects [29, 69, 65].

There is a model developed to predict DDIs based on the Interaction Profile Fingerprint (IPF) [56]. Quite simply, the interaction probability matrix was computed by multiplying the DDI matrix by the IPF matrix. Afterward, [40] proposed a computational framework by applying matrix perturbation, based on the hypothesis that by randomly removing edges from the DDI network, the eigenvectors of the adjacency matrix of the network should not change significantly. These two methods employ no other data about drugs, except known DDIs.

A new family of similarity-driven methods has followed the assumption that similar drugs should have almost similar interactions. Vilar et al. [19] presented a neighbor recommender method by utilizing substructure similarity of drugs. Relying on Vilar’s framework, Zhang et al. constructed a weighted similarity network that is labeled based on interaction with each of the drugs [67] and applied an integrative label propagation method using a random walk model on the network to estimate potential DDIs. This prediction framework only considered three types of similarities for predicting DDI via label propagation, namely substructure-based, side effect-based, and offside effect-based label propagation models [67]. Recently, some methods have also been proposed for adverse event detection using signals from social media [3, 4, 22].

In recent years, deep learning is becoming a promising technique for automatically capturing chemical compound features from data sets, and it successfully improves predictive performance. For example, Harada et al. [51] constructed a dual graph convolutional neural network to predict DDIs by combining the internal and external graph structures of drugs to learn low-dimensional representations of compounds.
However, this method works well only for moderately dense chemical networks with heavy-tailed degree distributions. Wang et al. [60] combined interview information of drug molecular and intraview of DDI relationships, developing a graph contrastive learning framework to predict DDIs. Lin et al. [39] merged several data sets into a vast knowledge graph with 1.2 billion triples, constructing KGNN to resolve the DDI prediction. On the other side, based on the structural, gene ontology term, and target gene similarity profiles, Lee et al. [37] applied an autoencoder to reduce the dimensions of each profile, constructing a DNN model by combining all the reduced features to predict the types of DDIs. Deng et al. [13] used the chemical substructures, targets, enzymes, and pathways of drugs to compute a similarity matrix of drugs, inputting each matrix to a DNN model, and combining the four submodels to predict DDI events. Besides DDI prediction, deep learning is also successfully applied for drug–target interaction prediction; for example, Shang et al. [53] develop a multilayer network representation learning method to learn the feature vectors of drugs and target. An et al. (An and Yu, [7]) use biased RWR and Word2vec algorithms to obtain the feature representation of drugs and targets. [63]

In this study, we develop a novel DDI prediction method utilizing the protein sequence data from the DrugBank and protein structure data from the Protein Data Bank. We calculate different similarity measures to create the similarity matrices for each feature attribute. Then, we use the generated feature matrices to create a single network fusion to measure the potential for interaction between two drugs. Final decision is performed via the help of a neural network architecture based on multi-layer perceptrons.

The main novelty of our approach is the focus on only genetic materials (protein sequence and protein structures) associated with the drug targets in developing our prediction model. To our knowledge, this is the first attempt at investigating potential DDI prediction by utilizing only information about the protein sequence and structure to generate the feature space fed to the neural network.
CHAPTER 3

Methodology

We developed a novel neural network model for the prediction of DDIs. The key idea in our approach is the notion that if two drugs have a similar pattern of similarity with other drugs, they are likely to have a similar pattern of interacting partners. To capture the patterns of similarity between drugs, we use information about the protein sequences and protein structures associated with the protein targets for a given drug.

Thus, we construct similarity matrices between drugs based on the protein sequences and protein secondary structures and combine these into one protein sequence-structure similarity matrix using network fusion. Fig. 3.4.1 shows a schematic diagram of the general proposed framework.

To calculate the similarity matrices we have used cosine distance, Levenshtein distance, Jensen Shannon (JS) divergence, and Euclidean Distance as the similarity measure between a pair of drugs.

3.1 Important Terminologies

Before starting about discussing the process of predicting the drug drug interaction, we will discuss about some of the basic and important terminologies to proper idea about the problem we are addressing here.
3. METHODOLOGY

3.1.1 Drug Protein targets

Drug protein targets are proteins found in the body of living animals that are linked to specific disorders for which medications are typically used to achieve the desired therapeutic effect. As a result, in order to be a pharmacological target protein, the protein must be linked to a disease process. Drug protein targets include enzymes, receptors, and transporter proteins, however receptors account for the majority of the targets. There are 1267 therapeutic protein targets that are pharmacologically active.

3.1.2 Drug Action

The molecular physiological mechanisms by which a chemical creates a response in living organisms are known as drug action. The alterations we observed after taking medications are referred to as drug action effects. Penicillin, for example, works by interfering with bacterial cell wall formation, resulting in the bacteria’s death. Drugs are mostly utilized to distinguish between normal metabolic processes and any anomalies that may exist. Because the differences may not be significant, drugs may work in a non-specific manner, altering both normal and unwanted processes. This results in unfavorable side effects.

3.1.3 Drug Pairwise Similarity vs Drug Action Similarity

Drug pairwise similarity is a mathematical representation of a relationship between two pharmaceuticals based on their protein information, which can be either sequence or structural information, with which two drugs would be similar if they were both used for the same ailment. We can construct similarity measures between two pharmaceuticals using several levels of information such as chemical structure, pharmacological targets, side-effects, indications, routes, and so on. We employed drug protein targets to calculate drug pairwise similarity in this paper.

Drug action similarity, on the other hand, is a measure of how similar two medications are in terms of how they affect living creatures. The general drug-cell interaction
can be used to determine the similarity of pharmacological activity. If two distinct pharmaceuticals have active implications on the same cell, we can say they have similar actions, but that doesn’t guarantee they will have the same type of effect on the human body.

### 3.1.4 Drug Similarity space

Drug Similarity space is a feature space of a list of pair of drugs with some attributes with which we can decide whether a pair of drugs will interact or not. If there is N number of drugs then we could have $N^2 - N$ possible pairs which might have interaction between them and might not.

### 3.1.5 K-mer

A k-mer is just a sequence of k characters in a string (or nucleotides in a DNA sequence). It is important to remember that to get all k-mers from a sequence you need to get the first k characters, then move just a single character for the start of the next k-mer and so on. Effectively, this will create sequences that overlap in k-1 positions.

Decomposing a sequence into its k-mers for analysis allows this set of fixed-size chunks to be analysed rather than the sequence, and this can be more efficient. K-mers are very useful in sequence matching (string matching with n-grams has a rich history), and set operations are faster, easier, and there are a lot of readily-available algorithms and techniques to work with them[9].

### 3.1.6 Suffix Array

A suffix array is a sorted array of all suffixes of a string. The definition is similar to Suffix Tree which is compressed trie of all suffixes of the given text. Any suffix tree-based algorithm can be replaced with an algorithm that uses a suffix array enhanced with additional information and solves the same problem in the same time complexity.
In short, the array of indexes to the sorted array of substrings generated during the transform is essentially a suffix array, which in turn is a representation of the information in a suffix tree [15].

### 3.2 Distance Matrices

To estimate the similarity between drugs, we compute distance measures (and sometimes similarity measures) between drugs based on their protein sequences and protein structure. In the context of data mining, a similarity measure is a distance with dimensions representing object features. When the distance between two items is little, they are quite similar, however when the distance is large, we will see a low degree of resemblance. There are several different types of similarity distance metrics. However, we’ll investigate Cosine Similarity, Levenshtein Distance, Jensen Shannon (JS) Divergence, and Euclidean Distance.

#### 3.2.1 Cosine Similarity (CS)

Cosine similarity metric finds the normalized dot product of two vectors. By determining the cosine similarity, we would effectively try to find the cosine of the angle between the two objects, when represented as vectors. The cosine of 0° is 1, and it is less than 1 for any other angle. For two $n$-length vectors $A$ and $B$, we have:

$$
CS(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}
$$

#### 3.2.2 Levenshtein Distance (L)

The Levenshtein distance is a string metric for measuring the difference between two sequences. The Levenshtein distance between two strings $a$, $b$ (of lengths $|a|$ and $|b|$, respectively) is given by $L_{a,b}(|a|, |b|)$
3. METHODOLOGY

\[ L_{a,b}(i, j) = \begin{cases} 
\max(i, j) & \text{if } \min(i, j) = 0 \\
L(i - 1, j) + 1 & \min \\
L(i, j - 1) + 1 & \text{otherwise} \\
L(i - 1, j - 1) + 1 & 
\end{cases} \] (2)

Essentially, \( L_{a,b}(i, j) \) is the distance between the first \( i \) character of \( a \) and the first \( j \) character of \( b \).

3.2.3 Jensen Shannon (JS) divergence (JSD)

The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. Given two distributions \( X \) and \( Y \), the JS divergence is the average KL divergence of \( X \) and \( Y \) from their mixture distribution, \( M \):

\[ JS(X||Y) = \frac{1}{2}D(X||M) + \frac{1}{2}D(Y||M) \] (3)

where \( M = \frac{X + Y}{2} \). and \( D(X||M) \) is the KL divergence between \( X \) and \( M \).

3.2.4 Euclidean Distance

This is the basic distance measure, defined as:

\[ ED(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \] (4)

When data is dense or continuous, this is the best proximity measure. The Euclidean distance between two points is the length of the path connecting them. This distance between two points is given by the Pythagorean theorem.

Here, we will also give a brief introduction of another distance metric name Tanimoto Coefficient which we use on other type of similarity networks on
3. METHODOLOGY

3.2.5 Tanimoto Coefficient

Tanimoto coefficient is determined by looking at the number of chemical features that are common to both molecules (the intersection of the data strings) compared to the number of chemical features that are in either (the union of the data strings). The Tanimoto coefficient is the ratio of the number of features common to molecules to the total number of features. i.e.

\[ Tanimoto = \frac{(A \cap B)}{(A + B - (A \cap B))} \] (5)

The range is 0 to 1 inclusive.

3.3 Similarity Matrices

In most of the DDI prediction methods, its very challenging to finding and developing the computational approaches which are appropriate drug features. For this study, we consider multiple data sources to collect different types of drug feature information to calculate the similarity matrices. In our work, we use similarity matrices, rather than distance matrices. Thus, for each distance measure, we convert the values into similarity measurement. In this section, we discuss the different types of matrices we calculated and used in our model evaluation.

3.3.1 Protein Sequence and Structure Similarity Matrices

Each protein structure could have multiple chains. Moreover, each drug active ingredient could have multiple protein targets. Thus, we could compute the similarity between two drugs (or drug active ingredients) based on the protein chains associated with the respective protein targets for the drugs. For each similarity measure, we record

1. Minimum Similarity
2. Maximum Similarity
3. METHODOLOGY

3. Average Similarity (AS)

4. Exponential Weighted Average Similarity (EWAS)

Here, we discuss briefly the protein sequence and protein structure similarity matrices used in this work.

3.3.1.1 Protein Sequence Similarity Matrices

In this approach, the protein sequence information is used directly to compute the similarity matrix. We can compute the Levenshtein distance directly. To compute the cosine, and JS divergence, we will first compute the $k$-mer profiles for each sequence, and then compute the similarity measure based on the profile. To generate the $k$-mer profiles, we use the suffix array data structure [15].

3.3.1.2 Protein Structure Similarity Matrices

For protein structure, we first covert the protein 3D structure into a protein string (pString) representation following [16]. The resulting pString is then treated like a sequence of information for structure. The only difference with the protein sequence is that each protein structure could have multiple chains (sequences) of information.

We generalized the similarity calculation which will represent the similarity values between two drug active ingredients (DAIs). We already know that each DAI could have multiple protein targets. Also, though each protein target has just one sequence, it could have multiple chains for its 3D structure. Thus, for a given DAI, we capture its protein structure information as follows:

$$[R_1^1, R_2^1, \ldots R_k^1, R_1^2, R_2^2, \ldots R_k^2, \ldots R_M^1, R_1^M, R_2^M, \ldots R_k^M]$$

where $R_i$'s represent the the protein targets, $M$ denotes the total number of protein targets in the DAI, and $k_1, k_2, \ldots, k_M$ represent the number of chains on each protein target.

Now, we can use this generalized DAI representation for similarity calculation between two DAI’s. If two DAI’s have $N$ and $M$ protein targets and their number
of chains are $k_1, k_2, ... k_N$ and $L_1, L_2, ... L_M$ respectively, then the possible number of
comparisons would be:

$$P_c = k_1L_1 + k_1L_2 + ... k_2L_1 + ... + k_ML_N$$ (6)

After $P_c$ comparisons at the chain level, we will obtain a vector of similarity values
for the two DAI's. We use the vector to calculate the minimum, maximum, average
and exponential weighted average similarity between the two DAI’s. The exponential
weighted average is computed as follows:

$$w_i = \frac{e^{s_i}}{\sum_i (e^{s_i})}$$ (7)

Here $w_i$ represents weights and $s_i$ represents a similarity value.

So, we used the minimum, maximum, average, and exponential weighted average
to calculate the similarities each time. As a result, when the similarity values were
determined, the following comparisons in equation 5 occurred on a chain level between
the two medications.

The total number of comparisons would be $K_1L_1 + K_1L2 + ... K_2L_1 + ... + K_ML_N.$

After getting all the quantitative analysis on protein chain level for protein struc-
ture data, We use the minimum,maximum,average and weighted average similarity
methods to calculate the similarity values between two drugs.

Each comparison will generate a similarity value $s$ which I will convert to $e^s$ So,
the chain level comparison will generate the similarity values below. So weight

$$w_i = \frac{e^{s_i}}{\sum e(e^{s_i})}$$ (8)

Then the weighted average similarity would be

$$w_is = w_1s_1 + w_2s_2 + ... + w_lsl$$ (9)

From the similarity values we will also have minimum similarity $mins$, maximum
similarities $max$s and average similarity as

\begin{align}
mins &= \min s_1, s_2, ... s_l \\
maxs &= \max s_1, s_2, ... s_l \\
as &= \frac{s_1, s_2, ... s_l}{l}
\end{align}

where $l$ is the number of pair of chains for two Protein structures, So, $l = M \times N$. After getting the minimum, maximum, average, weighted average from the chain level, we will now consider drug active ingredients (DAI) level min, max, avg, weighted average, avg min and avg max. This calculation is normal mathematical min, max and avg.

### 3.4 Protein Sequence-Structure Similarity Network (PS3N)

In Fig 3.4.1, Left hand side of the diagram represents the process of similarity metrics generation. Using the protein sequence, and protein structure similarity matrices, we generate protein sequence based, and protein structured based similarity networks using similarity network fusion approach. Each of these networks can be used independently to analyze potential DDIs between drugs or drug active ingredients. To improve the overall performance, we then integrate the sequence-based similarity network with the protein-based network into one overall similarity network. The result is the protein sequence-structure similarity network (PS3N). Our network integration is based on the technique of Similarity Network Fusion (SNF) [21]. SNF is an approach for combining multiple data sources into a single graph representing sample relationships.

The $k$-nearest neighbors approach is used in the similarity network construction and fusion process to down-weight weaker associations between samples. However,
3. METHODOLOGY

Fig. 3.4.1: Proposed Protein Sequence-Structure Similarity Network (PS3N) model for predicting adverse drug events. Using the method of Similarity Network Fusion (SNF) we create a single $N \times N$ fusion matrix for $N$ drugs. From the fusion matrix, we compute the feature vectors for each pair of drugs. In this way we will have possible $\binom{N}{2}$ rows, and each row will have $N$ columns as features. These feature vectors are then fed into a multi-layer perceptron model. For protein sequence similarity network, the number of hidden layers would reduce to 3 since we have less number of drugs.

weak relationships that are consistent across data sources are retained during the fusion process. The generated integrated network forms the basis for our analysis of adverse drug events, such as drug-drug interactions.

3.5 Neural Network Model

The model we propose for our problem is entirely reliant on the datasets we’re working with. That means the neural network’s performance will be influenced by the number of medications used in the dataset. In our neural network model, we used no more than four hidden layers. We use Rectified Linear activation function (ReLU) as the activation function where the dropout rate for each layer would vary from 0.3 to 0.5. Each of the hidden layers is followed by a dropout layer to avoid over-fitting problems during the training of the model. The output of each neuron in a layer is a nonlinear function $f$ of all nodes in the previous layer. $f$ is the ReLU, which is defined as the positive part of its arguments,

$$f(x) = x^* = \max\{x, 0\} \quad (13)$$
The final output layer is calculated using the sigmoid function:

\[ Sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (14) \]

For each layer, we used Xavier weight initialization, with Categorical cross-entropy as the loss function which also known as softmax loss. Its a softmax activation plus a cross-Entropy loss.

Though its normally used for multi-class classification, we used this to maintain the generality of our prediction. In multi-class classification the labels are one-hot, so only positive class \( C_p \) keeps its term in the loss. There is one element in the target vector \( t \) which is not zero \( t_i = t_p \). So discarding the elements of the summation which are zero due to target labels[27], we can write

\[ CE = -\log\left( \frac{e^{S_p}}{\sum_j e^{S_j}} \right) \quad (15) \]

Where, \( S_p \) is the CNN score for the positive class.

and Adam Optimizer is used as the optimizer in which a learning rate is maintained for each network weight (parameter) and separately adapted as learning folds. It combines momentum and Root Mean Square propagation (RMSProp) process to seep up the learning process. If \( m, v \) represents the momentum vector and \( \beta_1, \beta_2 \) as the exponential decay the update rules of Adam would be [14],

\[ \hat{m}^{k+1} = m^{k+1} \frac{1}{1 - \beta_1^{k+1}} \quad (16) \]
\[ \hat{v}^{k+1} = v^{k+1} \frac{1}{1 - \beta_2^{k+1}} \quad (17) \]

which will finally represents the optimization function as

\[ \theta^{k+1} = \theta^k - \eta \frac{\hat{m}^{k+1}}{\sqrt{\hat{v}^{k+1}} + \epsilon} \quad (18) \]
Cardio-oncology is a rapidly evolved field in which patients with cancer with cancer and cardiovascular disease are exposed to complex medication regimens, placing individuals at increased risk for potential drug interactions (DI’s)[11]. Most cardiovascular and cancer therapies tend to have complex pharmacological profiles, including intrapatient and inter-patient variability, narrow therapeutic index, and a steep dose-toxicity curve[11]. Annually, 17.9 millions deaths are caused by cardiovascular disease making 31% of all-cause mortality[24]. So patients with cardiovascular diseases (CVD)n are at high risk of experiencing drug-drug interactions. Because of multiple etiologies and concurrent comorbidities, CVD patients are treated with complex therapeutic regimen comprising multiple drugs[5].

As we already have the PS3N model to predict the DDI’s, We could utilize the model here for a cardiovascular disease problem. So far, we know how to create similarity networks. But from Drug Bank, Uniprot, SIDER, and other data sources, there are lot of information available with which more features can be generated.

Here, we will discuss other Drug Similarity profiles calculated based on different distance metrics and methodologies.
4. APPLICATION TO CARDIOVASCULAR DRUGS

4.1 Similarity Metrics based on other drug Information

When we create the PS3N model, we only considered protein sequence and protein structure information to predict the new DDI. However, there is a lot of other important information that have been used by several state-of-the-art algorithms. Though the way they utilize different drug features were different, that information’s are very important for drug interaction predictions. We utilize protein-protein interaction, side-effects, pathways, protein alignments, drug indication, and chemical Structure information to create a different type of drug similarity profile. Here we briefly discuss the different types of drug profiling or similarity network generation.

4.1.1 Protein-Protein Interaction (PPI) Network

We use HIPPIE(http://cbdm-01.zdv.uni-mainz.de/ mschaefer/hippie/) data source to collect the protein-protein interaction data. According to the data, we will have different PPIs with proteins ids and sequence information. What we did here is to we consider each interaction between pair of proteins as an edge and store them in a list. We also find out the unique IDs of proteins here. So from the PPI, we have two usable data structures,

- Edge List(PPI’s)
- Node list

From the edge list and the node list, we will create a graph which would be a network of protein interactions. We will utilize all possible pairs of the shortest path calculation technique to calculate the distances of possible all pairs of proteins. As we will consider each edge as bidirectional in the graph so if we calculate the distance between a pair of proteins for example A to B it would be similar for B to A. So, In this way, we will generate a matrix that represents the minimum distance from any node to all other nodes in the graph.
4. APPLICATION TO CARDIOVASCULAR DRUGS

The distance between two proteins should be 0 if there is no potential edge between them. The two proteins will interact directly if their minimal distance is 1. If the distance between two proteins is more than one, there will be an indirect contact between them. We are calling it a hop distance.

After creating the matrix of the protein-protein interaction network, we will utilize the pathway information to create the mapping from drugs in the pathways to the PPI matrix.

For each PP interaction, there is a value from the matrix. Suppose one value is D then consider the following equation as the converter of the hop distance of matrix into a similarity value S.

\[ S_{i,j} = e^{-D_{i,j}^2} \]  

where \( D_{i,j} \) = hop distance between the pair of proteins \( i \) and \( j \). So, after getting similarity value for each pair, I adjust the minimum, maximum and average for the drug-drug pair as we calculated for the PS3N.

4.1.2 Pathway Network

For pathway information, we calculated drug-drug edge frequency and protein-protein edge frequency information from the pathways. And we have a pathway protein distance matrix from which we calculate the distances of direct or hop-based interactions.

For calculating the drug-drug distances using the matrix, we will have a unique list of drugs from the drug information of pathways. As each drug has multiple proteins related to them, they will have like \((M \times N)\) (\( M = \) number of proteins for the first drug and \( N = \) number of proteins for the second drug) pair of distances from the matrix. So, from the \((M \times N)\) pair we will calculate the minimum, maximum, and average similarity. We will also calculate the weighted similarity between drugs.

As the distance matrix contain the hop distance information, we used the following formula to convert them into similarity.

\[ S_{i,j} = e^{-D_{i,j}^2} \]  

\[ (2) \]
where $S_{i,j} = \text{Similarity between a protein pair } i \text{ and } j$. $D_{i,j} = \text{hop distance between the pair of proteins } i \text{ and } j$. So, after getting similarity value for each pair, we adjust the minimum, maximum and average for the drug-drug pair.

### 4.1.2.1 Weighted Similarity of Drug pathway Network

For weighted similarity, we had the protein-protein-edge frequency information but how we utilize the data in our weighted similarity calculation is a big challenge. Because till now we only have the frequency for those edges who have the direct interactions. So, for those who have indirect interaction we need to formulate a way to find the frequencies so that our weighted similarity would be describable.

Suppose we have two drugs A and B, A has 3 proteins E,F,G and B has 2 proteins P,Q we will have $3 \times 2 = 6$ protein pairs $(E,P),(E,Q),(F,P),(F,Q),(G,P),(G,Q)$

If any pair have direct edge then we will get the frequency from the protein-protein edge frequency data. But when we don’t have the direct interaction. We consider the shortest path between them. Suppose between E,Q, we have a path: $E \rightarrow s_1 \rightarrow s_2 \rightarrow Q$

There might be other paths as well, but we focus on the shortest path for the interaction of E and Q. So, to calculate the frequency between E and Q we might consider the frequency median of $E \rightarrow s_1$, $s_1 \rightarrow s_2$, and $s_2 \rightarrow Q$. The median would be the edge frequency for $(E,Q)$ proteins edge. Now suppose the newly calculated frequency is $f_1$. Then the equation of weighted similarity would be

$$WS_{AB} = \sum_{i \in A, j \in B} \frac{S_{i,j}w_{i,j}}{MN}$$

(3)

where $w_{i,j}$ represents the weights and $i,j$ represents the similarity value based on the hop distance.

In this way we will generate all the similarities for the drug pairs. Directly using the drug-drug edge information from the pathway to calculate the distance matrix. Same as the previously mentioned approach.
4. APPLICATION TO CARDIOVASCULAR DRUGS

4.1.3 Protein-Protein Alignment Similarity Network

In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages.

There are two types of alignments 1. Global 2. Local alignments. Here we mainly consider global alignment to generate the similarity network. PAM250 matrix[58] is frequently used to score aligned peptide sequences to determine the similarity of those sequences. The numbers given above were derived from comparing aligned sequences of proteins with known homology and determining the accepted point mutation. The frequencies of these mutations are tabular form as a log odds-matrix where

\[ M_{ij} = 10(\log_{10} R_{ij}) \] (4)

Where \( M_{ij} \) is the matrix element and \( R_{ij} \) is the probability of that substitution as observed in the database, divided by the normalized frequency of occurrence for amino acid \( i \). All the number are rounded to the nearest integer. The base-10 log is used so that the numbers can be added to determine the score of a compared set of sequences, rather than multiplied.

Using the protein target sequences, we’ll create alignments for paired protein target sequences using the PAM250 matrix. The alignment score and align length, as well as the aligned sequences for the pairs, will be returned by the function. We’ll use the Open Gap and Extend approach to calculate the score. When we find an amino acid match in a sequence, we assign it an identity score; if we find a mismatch, we open a gap with a penalty value and expand it with an extending penalty value. It’d always be a negative number.

We’ll count the number of identity matches from the sequences after we’ve found the two aligned sequences. To compute the similarity percentage between two sequences, divide the aligned length by the identity score. The ratio of the number of matching residues to the overall length of the alignment is the similarity percentage.
4.1.4 Side-Effects Similarity Network

We use the cosine Similarity approach to calculate the similarity value of DAI s in order to establish a similarity network based on side-effect data. We started by constructing a list of acronyms for side effects from the dataset. We used a word-to-vector method to turn the word information into a vector. We are aware of a variety of feature extraction methods, such as Bag of Words (BOW) and TF-IDF, among others. BOW is the most straightforward and intuitive method, involving two types of calculations. One method is to count the terms in the list of side effect names. The other is to figure out how often each word appears. The BOW model, on the other hand, has a problem in that words with a higher frequency dominate word lists, even if they are unrelated to the other words in the list. That is why we use the TF-IDF approach in this case. It rescales a word’s frequency based on how many times it appears across all word lists. Words that are frequently used, such as ”the,” ”that,” and others, receive a lower score and are penalized.

TF-IDF score of a word $w = tf(w) \times idf(w)$ where

$$tf(w) = \frac{\text{Number of times the word appears in the list}}{\text{Total number of words in the list}}$$

(5)

and

$$idf(w) = \frac{\text{Number of drugs}}{\text{Number of drugs that contains word w}}$$

(6)

Here, we can see the tf-idf numerical vectors contains the score of each of the words of side effects in the DAI. So, in this way we will convert the side effects of a given DAI into numerical features using TF-IDF. We will use cosine similarity function to compare the first DAI with other DAI in the corpus. The generated result would be a matrix showing the similarity of each drug to every other drugs.

4.1.5 Chemical Structure Drug Similarity Network

Molecular fingerprints, which record structural information about a molecule as a series of bits, are commonly used to compute the similarity between two molecules.
These bits signify the presence or absence of specific patterns or substructures; two molecules with more of the same patterns will share more bits, suggesting that they are more similar.

The Pubchempy library has Compound information of each PubChem ID. The PubChem CACTVS fingerprint is available on each compound using the fingerprint method. The information I provided about the two DAIs is hexadecimal encoded representation. We will decode this from hexadecimal and convert them as binary.

Then we will calculate the similarity based on the binary representation of the molecular fingerprints using the tanimoto coefficient method.

4.2 Protein Sequence, Structure and Chemical Structure Similarity Network (PSSCSSN)

To create the PS3N network, we previously fused protein sequence and protein structure matrices. To develop the final network, we analyzed all forms of similarity matrices to predict drug interactions in cardiovascular disease. However, there is a serious issue with data divergence. It was challenging to obtain all forms of information for a medicine because most types of data were acquired from several data sources. After a thorough examination of all accessible datasets, we discovered that the Subset of Protein Sequence, Protein Structure, and Chemical Structures has the most pharmaceuticals for which our neural network model can be trained. We found 722 medicines with all three pieces of information thanks to the interaction of the three datasets.

So, we consider the Similarity matrices from Chemical Structures, protein Structures, and Protein sequences in SNF[21] approach. By Applying SNF, we finally get our desired PSSCSSN Network.
4. APPLICATION TO CARDIOVASCULAR DRUGS

4.3 Neural Network Model

To train the PSSCSSN network for predicting drug-drug interactions in inter cardiovascular drugs and interactions with non-cardiovascular drugs, we have to consider a neural network model which would take a subset of the network during the training.

This neural network model, like the PS3N model, will be dependent on the amount of medicines. As a result, the input layer will have 722 Neurons, while the first concealed levels will have 490 neurons. To avoid the over-fitting problem during training, there will be a dropout layer with a comparable dropout rate as before. As the activation layer, the ReLU was utilized. The number of layers increases to 300 in the second concealed layer. The third concealed layer will contain 150 neurons, while the fourth layer will have 50 neurons. The loss function in this model is Categorical Cross Entropy. The Adam optimizer is the optimization function that begins training with the default learning rate. The Sigmoid function is used as the output layer of the model.
CHAPTER 5

Experiments and Results

For the development of the PS3N model to predict effectively the DDI’s, we put significant time on data processing and cleaning. Our whole experimental setup is divided into multiple parts. Before going to the experiments and results here is a brief introductions to the datasets we utilize to prepare our similarity matrices.

5.1 Datasets

In this work, we use different sources for different types of drug information. Most of the information’s like Protein Sequence, Pathways are extracted from Drug Bank (https://go.drugbank.com/).

The second is the protein structure dataset retrieved from RCSB Protein Data Bank (https://www.rcsb.org). Protein chains are extracted from each PDB file using biopython libraries. We combined these for a dataset of 905 drugs (active ingredients) in DrugBank with information on both protein structure and protein sequences. The side-effects and indications information of drugs were collected from SIDER (http://sideeffects.embl.de/) database. Protein-Protein Interaction information were collected from Uniprot (https://www.uniprot.org/uniprot/) portal. Lastly, We also evaluated our methods using the DS1 and DS2 datasets reported in [41].
5.2 Performance Evaluation of PS3N Model

To evaluate the performance of the proposed method, we compared it with machine learning approaches such as KNN (K-Nearest Neighbor), RF (Random Forest), Logistic Regression, LDA (Linear Discriminant Analysis), and Support Vector Machine. We also compared our results with state of the art methods proposed in [41], [56], [67], [68]. We evaluated the competitiveness of our models using different performance metrics such as Precision, Recall, F1, Area under Curve (AUC), and AUPR. Here, I am briefly providing the calculation approach for each of the performance metrics.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (1)
\]

\[
\text{recall} = \frac{TP}{TP + FN} \quad (2)
\]

\[
F - \text{measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)
\]

Where TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative. Precision is the fraction of correct predicted interactions among all predicted interactions, while recall is the fraction of correct predicted interactions among all true interactions. Precision and recall have a trade-off thus improving one of them may lead to a reduction in another. Therefore, utilizing F-measure which is the geometric mean of precision and recall is more reasonable.

We note that if the interaction between two drugs is assigned to zero, it simply implies that no evidence of their interaction has been found yet. The two may still interact, but the features we have used so far are not able to detect that.

Thus, they may interact with each other. So we cannot identify TN and FP pairs correctly. The training process requires both positive and negative samples. Therefore, some of the zero assigned pairs are considered as non-interactive pairs in the training model. So every method may have some FP in its evaluations. This leads to a reduction in calculated precision and F-measure, while the real values of precision and F-measure may be higher. Since the values of precision, recall, and
F-measure is dependent to the value of the threshold, we also evaluate methods via AUC which is the area under the receiver operating characteristic (ROC) curve.

We note that if the interaction between two drugs is assigned to zero, it simply implies that no evidence of their interaction has been found yet. The two may still interact, but the features we have used so far are not able to detect that.

5.2.1 Experimental setups

For our training experiment, we split each dataset into training, validation, and test sets according to a 70% − 10% − 20% random split. For each dataset, networks were trained on the training set for a total of 100 epochs with a batch size of 100 for the proposed neural network method.

To initialize the weights of the network so that the neuron activation function can avoid saturation problems or being stuck into dead regions. We used batch size of 100 and 20 - 50 epochs, with Categorical cross entropy and Adam Optimizer for optimization with a momentum parameter 0.9. The epoch number was set to 20 and 50. We use Categorical cross entropy and Adam optimizer for optimization with a momentum parameter 0.9.

5.2.2 Comparing performance with State-of-the-art Algorithms

We evaluated our model on the single feature matrices to identify the contribution of specific features to the performance of the model. We used the average and exponential weighted average similarity measures to generate the similarity matrices. Table 5.2.1 shows the performance of the proposed model using protein sequences. Table 5.2.2 shows results using protein structure.

In Table 5.2.4 and 5.2.5 we compare our results with the existing state-of-the-art algorithms and found significant improvements in terms of AUC, Precision, and Recall. We considered DS1 and DS2 datasets from [41] to compare the performance of existing methodologies. From the two datasets, we could generate the protein sequence and structure metrics for a subset of drugs. We used the newly generated
Table 5.2.1: Performance of PS3N using similarity matrices based on protein sequences

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>L AS</td>
<td>0.9199</td>
<td>0.9419</td>
<td>0.9308</td>
<td>0.9673</td>
<td>0.9081</td>
</tr>
<tr>
<td>JSD AS</td>
<td>0.8837</td>
<td>0.8667</td>
<td>0.8751</td>
<td>0.9181</td>
<td>0.8377</td>
</tr>
<tr>
<td>CS AS</td>
<td>0.8799</td>
<td>0.9093</td>
<td>0.8943</td>
<td>0.9315</td>
<td>0.8590</td>
</tr>
<tr>
<td>L EWAS</td>
<td>0.9499</td>
<td>0.9638</td>
<td>0.9568</td>
<td>0.9832</td>
<td>0.9429</td>
</tr>
<tr>
<td>JSD EWAS</td>
<td>0.9406</td>
<td>0.8835</td>
<td>0.9112</td>
<td>0.9559</td>
<td>0.8870</td>
</tr>
<tr>
<td>CS EWAS</td>
<td>0.9523</td>
<td>0.9632</td>
<td>0.9578</td>
<td>0.9833</td>
<td>0.9443</td>
</tr>
</tbody>
</table>

L = Levenshtein, JSD = JS Divergence, CS = Cosine, AS = Average Similarity, EWAS = Exponential Weighted Average Similarity

feature space in our model to check the performance and showed significant improvement in both cases. From Table 5.2.4, we can see that the PS3N showed better performance when compared to the other methods. However, it showed similar results on the datasets based on sequence, structure, or both information. This also holds in Table 5.2.5 which was created from the DS2 dataset.

5.2.3 Predicted DDI Network Diagram Using PS3N

We quantified all the drug-drug interactions using the PS3N model in Fig. ?? As a result, when we anticipate drug interactions between pairs of pharmaceuticals, the model generates a likelihood of interaction between the drugs, which we refer to as an interaction score, which ranges from 0 to 1. We can also use the model to get the projected labels for the test data. To prepare network data, the pharmaceutical Ids are employed as network nodes. We considered a pair of drugs to have an advantage if the expected labels from the model were 1. We discovered that the network data contained a wealth of interaction information. We proposed imposing a threshold on the principal Similarity network to eliminate some data from consideration to improve
Table 5.2.2: Performance of PS3N using similarity matrices based on protein structure.

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>JSD AS</td>
<td>0.8796</td>
<td>0.9279</td>
<td>0.90313</td>
<td>0.9171</td>
<td>0.8564</td>
</tr>
<tr>
<td>CS AS</td>
<td>0.9037</td>
<td>0.9469</td>
<td>0.9248</td>
<td>0.9499</td>
<td>0.8889</td>
</tr>
<tr>
<td>JSD EWAS</td>
<td>0.9762</td>
<td>0.9650</td>
<td>0.9706</td>
<td>0.9895</td>
<td>0.9578</td>
</tr>
<tr>
<td>CS EWAS</td>
<td>0.9743</td>
<td>0.9738</td>
<td>0.9741</td>
<td>0.9910</td>
<td>0.9627</td>
</tr>
</tbody>
</table>

*L = Levenshtein, JSD = JS Divergence, CS = Cosine, AS = Average Similarity, EWAS = Exponential Weighted Average Similarity

Table 5.2.3: Results using combined protein sequence and protein structure similarity matrices.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS3N</td>
<td>0.9800</td>
<td>0.9818</td>
<td>0.9809</td>
<td>0.9946</td>
<td>0.9725</td>
</tr>
<tr>
<td>RF</td>
<td>0.7812</td>
<td>0.7241</td>
<td>0.7516</td>
<td>0.8089</td>
<td>0.8354</td>
</tr>
<tr>
<td>SVM</td>
<td>0.5507</td>
<td>0.2076</td>
<td>0.3015</td>
<td>0.5711</td>
<td>0.7320</td>
</tr>
<tr>
<td>LR</td>
<td>0.5169</td>
<td>0.1586</td>
<td>0.2427</td>
<td>0.5506</td>
<td>0.7243</td>
</tr>
<tr>
<td>LDA</td>
<td>0.5302</td>
<td>0.1740</td>
<td>0.2620</td>
<td>0.5572</td>
<td>0.7269</td>
</tr>
<tr>
<td>KNN</td>
<td>0.5470</td>
<td>0.6196</td>
<td>0.5810</td>
<td>0.7107</td>
<td>0.7510</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.7134</td>
<td>0.7008</td>
<td>0.7071</td>
<td>0.7961</td>
<td>0.8382</td>
</tr>
<tr>
<td>NDD</td>
<td>0.5646</td>
<td>0.1927</td>
<td>0.2874</td>
<td>0.7366</td>
<td>0.7311</td>
</tr>
</tbody>
</table>
Table 5.2.4: Performance comparison of different methods on DS1 from [41]. We obtained information on 469 drugs for protein sequences, and on 414 drugs for protein structure. The first six rows are from [41] to compare the results from our model.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>AUPR</th>
<th>F-measure</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure-based label propagation model [67]</td>
<td>0.937</td>
<td>0.901</td>
<td>0.804</td>
<td>0.797</td>
<td>0.811</td>
</tr>
<tr>
<td>Side-effect-based label propagation model [67]</td>
<td>0.936</td>
<td>0.903</td>
<td>0.806</td>
<td>0.793</td>
<td>0.820</td>
</tr>
<tr>
<td>Offside-effect-based label propagation model [67]</td>
<td>0.937</td>
<td>0.904</td>
<td>0.809</td>
<td>0.795</td>
<td>0.823</td>
</tr>
<tr>
<td>Vilar’s substructure-based model [56]</td>
<td>0.936</td>
<td>0.902</td>
<td>0.804</td>
<td>0.797</td>
<td>0.812</td>
</tr>
<tr>
<td>Classifier ensemble method [68]</td>
<td>0.956</td>
<td>0.928</td>
<td>0.836</td>
<td>0.827</td>
<td>0.843</td>
</tr>
<tr>
<td>Weighted average ensemble method [68]</td>
<td>0.948</td>
<td>0.919</td>
<td>0.831</td>
<td>0.835</td>
<td>0.826</td>
</tr>
<tr>
<td>NDD [41]</td>
<td>0.954</td>
<td>0.922</td>
<td>0.835</td>
<td>0.836</td>
<td>0.833</td>
</tr>
<tr>
<td>PS3N (Protein Sequence)</td>
<td><strong>0.974</strong></td>
<td><strong>0.948</strong></td>
<td><strong>0.916</strong></td>
<td>0.925</td>
<td><strong>0.906</strong></td>
</tr>
<tr>
<td>PS3N (Protein Structure)</td>
<td>0.972</td>
<td><strong>0.949</strong></td>
<td>0.917</td>
<td><strong>0.932</strong></td>
<td>0.903</td>
</tr>
<tr>
<td>PS3N (Sequence + Structure)</td>
<td>0.972</td>
<td>0.948</td>
<td>0.917</td>
<td>0.931</td>
<td>0.903</td>
</tr>
</tbody>
</table>

Presentation. We can see that several interactions have been discovered in Fig. ??, but they are all the same shade. Our primary goal was to see if the model could do DDI prediction with the maximum level of accuracy. When we applied the threshold to the network’s similarity values, we discovered that 89 percent of the interactions have the correct labels. When we looked at the entire network, however, performance rose by about 96 percent. Even if some of the estimated values have a similarity of less than 0.70, they nonetheless had interactions.

### 5.3 Impact of algorithmic parameters

Table 5.3.1 shows the impact of different hyperparameters on the performance of the proposed model. From the table, Adam Optimizer with a learning rate of 0.01 produced the best overall result. SGD Optimizer for learning rate 0.05, 0.01 and 0.10 showed almost similar accuracy level as we got for Adam optimizer. In our proposed neural network model, the number of hidden layers will vary based on the number of drug active ingredients (DAI’s) on the datasets. Normally, for protein sequence
Table 5.2.5: Performance comparison of different methods on the DS2 Dataset from [41]. We obtained information on 585 drugs for protein sequences, and on 504 drugs for protein structure. The first Six rows are from [41] to compare the results from our model.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>AUPR</th>
<th>F-measure</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure-based label propagation model [67]</td>
<td>0.788</td>
<td>0.208</td>
<td>0.294</td>
<td>0.537</td>
<td>0.197</td>
</tr>
<tr>
<td>Vilar’s substructure-based model [56]</td>
<td>0.810</td>
<td>0.244</td>
<td>0.312</td>
<td>0.479</td>
<td>0.232</td>
</tr>
<tr>
<td>Classifier ensemble method [68]</td>
<td>0.936</td>
<td>0.487</td>
<td>0.553</td>
<td>0.689</td>
<td>0.462</td>
</tr>
<tr>
<td>Weighted average ensemble method [68]</td>
<td>0.646</td>
<td>0.440</td>
<td>0.15</td>
<td>0.226</td>
<td>0.118</td>
</tr>
<tr>
<td>NDD [41]</td>
<td>0.994</td>
<td>0.890</td>
<td>0.825</td>
<td>0.804</td>
<td>0.847</td>
</tr>
<tr>
<td>PS3N (Protein Sequence)</td>
<td>0.998</td>
<td>0.975</td>
<td>0.978</td>
<td>0.987</td>
<td>0.972</td>
</tr>
<tr>
<td>PS3N (Protein Structure)</td>
<td>0.997</td>
<td>0.975</td>
<td>0.978</td>
<td>0.992</td>
<td>0.964</td>
</tr>
<tr>
<td>PS3N (Sequence + Structure)</td>
<td>0.997</td>
<td>0.970</td>
<td>0.977</td>
<td>0.987</td>
<td>0.970</td>
</tr>
</tbody>
</table>

In our proposed approach, similarity network fusion (SNF) used for final matrix creation of each datasets. We had different distance measures, for each of the distances we had one matrix. From the matrices we create the fusion matrix. During the creation of the network, some parameters were used which played a role to the similarity matrix construction. In the SNF method, a distance metric was considered, we used the default euclidean distance metric. And there were two other parameters $\mu$-weighted $k$-nearest neighbors kernel to the distance matrix to calculate the similarity fusion matrix. We used different values in range of 0.3 to 0.5 for $\mu$ and 10 to 20 for $k$. we found that for $\mu = 0.3$ and $k = 10$ works better for the performance of the model. So the generated fusion network with the parameter setup works better than the others.

In the deep neural network model, we used 3 to 5 hidden layers based on the number of drug active ingredients presents on the datasets. However in our proposed PS3N model we used he normal distribution for the layer initialization when we used ReLU as the activation function in the layer. We used he normal because we used
5. EXPERIMENTS AND RESULTS

Fig. 5.2.1: Drug Drug Interaction Similarity Network Diagram. We have used a threshold value $\geq 0.7$ in PS3N Network to discard a large chunk of network diagram. If any pair of drugs have predicted labels as 1, we consider an edge between the drugs. In these way we created a Network diagram for PS3N Model using PS3N Network.

Table 5.3.1: Results of PS3N with variation on algorithmic parameters.

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>Learning rate</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam optimizer</td>
<td>0.05</td>
<td>0.7200</td>
</tr>
<tr>
<td>Adam optimizer</td>
<td>0.10</td>
<td>0.7213</td>
</tr>
<tr>
<td>Adam optimizer</td>
<td>0.01</td>
<td>0.9710</td>
</tr>
<tr>
<td>SGD</td>
<td>0.05</td>
<td>0.9646</td>
</tr>
<tr>
<td>SGD</td>
<td>0.10</td>
<td>0.9644</td>
</tr>
<tr>
<td>SGD</td>
<td>0.01</td>
<td>0.9637</td>
</tr>
<tr>
<td>RMSProp</td>
<td>0.01</td>
<td>0.7210</td>
</tr>
</tbody>
</table>
5. EXPERIMENTS AND RESULTS

ReLU as activation. In the final layer where we used glorot normal distribution for the layer initialization because Sigmoid function used as the activation function. Though different initialization have very little impact on the performance, we select this as general set up for the experiment.

Then we used the dropout value with a range of 0.2 to 0.5 for each of the layer. 0.3 showed best outcome in the model performance. Then in the model optimization we used different optimization techniques like Stochastic gradient descent (SGD), Adam optimizer, RMSProp etc in which Adam and SGD have close performance but in RMSProp the algorithm performance decreases significantly.

In gradient calculation, learning rate is very important. In our experimental setup we choose different learning rate in a range from 0.001 to 0.1. We also use weight decay for the gradient update in a range from $1e^{-4}$ to $1e^{-8}$. But we found out the best value of weight decay for the model is $1e^{-6}$. After different selection of learning rate we found that 0.01 give us the best result for the model.

5.4 Results on Cardiovascular Drugs

To check the performance of the PSSCSSN model, we compared this new model with our PS3N. In Table 5.4.1, we can see that there are different PS3N variations. Basically, all the variations are representing different data sets.

In this experimental setup, we first considered evaluating the performance of the PS3N model on different datasets. Though the PS3N model is based on protein sequence and protein structure similarity network. We utilize the neural network model for other data sets like side-effects, pathways, protein-protein interactions, and chemical structures. To keep in mind the dependency on the drug numbers we have to adjust the input layer and hidden layer numbers for side effects and chemical structure data. there is 936 drugs’ side-effect information available from which the side-effect similarity network was calculated. On the other hand, the similarity network for Chemical structure has 1461 drugs into consideration. So the two different networks had a little different architecture in terms of hidden layers, but maintain the core
Table 5.4.1: Performance comparison of PS3N model and PSSCSSN model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS3N (pathways)</td>
<td>0.923</td>
<td>0.96</td>
<td>0.942</td>
<td>0.974</td>
<td>0.922</td>
</tr>
<tr>
<td>PS3N(Side-Effects)</td>
<td>0.895</td>
<td>0.921</td>
<td>0.908</td>
<td>0.951</td>
<td>0.883</td>
</tr>
<tr>
<td>PS3N(Chemical Structure)</td>
<td>0.964</td>
<td>0.983</td>
<td>0.974</td>
<td>0.991</td>
<td>0.964</td>
</tr>
<tr>
<td>PS3N (Protein Sequence)</td>
<td>0.998</td>
<td>0.975</td>
<td>0.978</td>
<td>0.987</td>
<td>0.972</td>
</tr>
<tr>
<td>PS3N (PPI)</td>
<td>0.958</td>
<td>0.967</td>
<td>0.962</td>
<td>0.979</td>
<td>0.954</td>
</tr>
<tr>
<td>PS3N (Protein Structure)</td>
<td>0.997</td>
<td>0.975</td>
<td>0.978</td>
<td>0.992</td>
<td>0.964</td>
</tr>
<tr>
<td>PS3N (Sequence + Structure)</td>
<td>0.997</td>
<td>0.970</td>
<td>0.977</td>
<td>0.987</td>
<td>0.970</td>
</tr>
<tr>
<td>PSSCSSN</td>
<td>0.981</td>
<td>0.970</td>
<td>0.975</td>
<td>0.974</td>
<td>0.970</td>
</tr>
</tbody>
</table>

design of the PS3N model. So we can call them variations of PS3N.

In 5.4.1 we can see that PS3N is performing better for most of the cases in terms of all performance metrics. If we consider accuracy metrics for the DDI prediction, PS3N (Sequence) is better than all other variations but we can say PSSCSSN and PS3N (Sequence + Structure) are pretty close in terms of accuracy metrics. However, we designed the PSSCSSN model only because we have a large subset of information from all the available datasets. Our main goal is to identify as much as possible to find the unknown DDI for the cardiovascular drugs and create a knowledge base from the given network information.

5.4.1 Predicted Interactions of cardiovascular drugs

Our model is building a database of pharmacological interactions, which we will use to better understand the systemic effects on cardiovascular disease. On the level of sequence and structure, the PSSCSSN model provides a different quantitative analysis of similarity values and their intended protein targets. Most importantly, we were able to predict the DDIs for both inter-cardiovascular medication interactions and cardiovascular-non-cardiovascular drug interactions using the model. Our projected
DDIs reveal a plethora of novel medication interactions.

We can see some quantitative analysis of the interactions in Table 5.4.2. We received a total of 2759 new DDIs from the PSSCSSN network. There are options for both types of interactions. Pravastatin, with a total of 6.38% new interactions, has the highest number of new interactions. Interactions with inter-cardiovascular medicines account for only 0.65% of those interactions.

For example, drugs like hydrocortamate and olanzapine interact with Pravastatin, causing dizziness, restlessness, depression, low blood pressure when standing, and other side effects. As a result, cardiovascular patients who take those medications for skin illnesses or mental health could be jeopardizing their overall health.

In Table 5.4.3, we can see a comprehensive representation of all the interactions found for inter cardiovascular DDI’s. The table represents a greater knowledge base for this CVD problem. There is a lot of information we can get from this table. It’s showing some surprising results when we look at the interactions between cardiovascular drugs and other drugs, because many common cardiovascular drugs interact with other CVD drugs, but there is no interaction between them. For example, if we look into the CVD drug Fenofibric Acid, we can see that there is only one
inter cardiovascular drug interaction with Candesartan. But when we looked into DDI's for other cardiovascular drugs we found Clopidogrel, Dronedarone, Ticagrelor, Donepezil, Benazepril,Dipyridamole, Olmesartan, Argatroban, Quinapril, Telmisartan, and Dofetilide interacting with Fenofibric Acid. There are several examples like this. So, this knowledge base could be vital to identifying the several relations among different drugs and their adverse reactions to the patients.

Table 5.4.3: Discovered new Inter cardiovascular DDI’s using PSSCSSN Model.

<table>
<thead>
<tr>
<th>Drug Name</th>
<th># Cardiovascular DDI’s</th>
<th>interacted cardiovascular drugs</th>
<th>% of cardiovascular DDI’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pravastatin</td>
<td>18</td>
<td>Ticlopidine; Milrinone; Disopyramide; Ibutilide; Diltiazem; Amlodipine; Triamterene; Benazepril; Fondaparinux; Eplerenone; Perindopril; Irbesartan; Terazosin; Hydralazine; Prasugrel; Azilsartan medoxomil; Aliskiren; Enalaprilat</td>
<td>0.65</td>
</tr>
<tr>
<td>Atorvastatin</td>
<td>9</td>
<td>Papaverine; Terazosin; Propafenone; Prasugrel; Azilsartan medoxomil; Aliskiren; Vorapaxar; Edoxaban; Enalaprilat</td>
<td>0.32</td>
</tr>
<tr>
<td>Drug</td>
<td>Code</td>
<td>Ingredients</td>
<td>Score</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Lovastatin</td>
<td>23</td>
<td>Milrinone; Disopyramide; Ibutilide; Amiodipine; Triamterene; Bendrofluamide; Trandolapril; Benazepril; Fonda-parinux; Niacin; Guanabenz; Moexipril; Perindopril; Quinapril; Dopamine; Hydrochlorothiazide; Guanfacine; Papaverine; Terazosin; Azilsartan medoxomil; Aliskiren; Edoxaban; Enalaprilat</td>
<td>0.83</td>
</tr>
<tr>
<td>Spironolactone</td>
<td>4</td>
<td>Niacin; Clofibrate; Dronedarone; Ticagrelor</td>
<td>0.14</td>
</tr>
<tr>
<td>Terazosin</td>
<td>3</td>
<td>Gemfibrozil; Prasugrel; Pitavastatin</td>
<td>0.10</td>
</tr>
<tr>
<td>Niacin</td>
<td>9</td>
<td>Guanabenz; Verapamil; Clopidogrel; Perindopril; Telmisartan; Dipyridamole; Guanfacine; Terazosin; Hydralazine; Prasugrel; Ticagrelor; Aliskiren</td>
<td>0.32</td>
</tr>
<tr>
<td>Cerivastatin</td>
<td>14</td>
<td>Fosinopril; Trandolapril; Benazepril; Guanabenz; Moexipril; Lisinopril; Perindopril; Dopamine; Guanfacine; Papaverine; Hydralazine; Ticagrelor; Azilsartan medoxomil; Aliskiren</td>
<td>0.51</td>
</tr>
<tr>
<td>Dofetilide</td>
<td>22</td>
<td>Milrinone; Trandolapril; Benazepril; Fonda-parinux; Enalapril; Guanabenz; Simvastatin; Perindopril; Donepezil; Quinapril; Pindolol; Telmisartan; Methyldopa; Guanfacine; Atorvastatin; Gemfibrozil; Prasugrel; Pitavastatin; Aliskiren; Sacubitril; Fenofibric acid; Candesartan</td>
<td>0.81</td>
</tr>
</tbody>
</table>
### 5. EXPERIMENTS AND RESULTS

<table>
<thead>
<tr>
<th>Drug</th>
<th>Value</th>
<th>Co-occurring Drugs</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ticlopidine</td>
<td>12</td>
<td>Olmesartan; Trandolapril; Benazepril; Enalapril; Moexipril; Eprosartan; Quinapril; Telmisartan; Papaverine; Captopril; Azilsartan medoxomil; Aliskiren</td>
<td>0.43</td>
</tr>
<tr>
<td>Clofibrate</td>
<td>17</td>
<td>Losartan; Moexipril; Perindopril; Donepezil; Quinapril; Telmisartan; Dipyridamole; Guanfacine; Irbesartan; Papaverine; Propafenone; Sitagliptin; Hydralazine; Azilsartan medoxomil; Aliskiren; Enalaprilat; Candesartan</td>
<td>0.62</td>
</tr>
<tr>
<td>Telmisartan</td>
<td>3</td>
<td>Fluvastatin; Ticagrelor; Fenofibric acid</td>
<td>0.10</td>
</tr>
<tr>
<td>Quinapril</td>
<td>4</td>
<td>Atorvastatin; Ticagrelor; Pitavastatin; Fenofibric acid</td>
<td>0.14</td>
</tr>
<tr>
<td>Argatroban</td>
<td>16</td>
<td>Ibutilide; Diltiazem; Fosinopril; Niacin; Losartan; Eplerenone; Perindopril; Acetazolamide; Telmisartan; Propafenone; Gemfibrozil; Azilsartan medoxomil; Pitavastatin; Aliskiren; Enalaprilat; Fenofibric acid</td>
<td>0.58</td>
</tr>
<tr>
<td>Fondaparinux</td>
<td>9</td>
<td>Guanabenz; Clofibrate; Donepezil; Telmisartan; Atorvastatin; Fluvastatin; Terazosin; Propafenone; Aliskiren</td>
<td>0.33</td>
</tr>
<tr>
<td>Olmesartan</td>
<td>10</td>
<td>Ibutilide; Clofibrate; Acetazolamide; Donepezil; Fluvastatin; Gemfibrozil; Sitagliptin; Prasugrel; Ticagrelor; Fenofibric acid</td>
<td>0.36</td>
</tr>
<tr>
<td>Drug</td>
<td>Frequency</td>
<td>Associated Drugs</td>
<td>Similarity Score</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Acetazolamide</td>
<td>12</td>
<td>Eprosartan; Telmisartan; Irbesartan; Fluvastatin; Papaverine; Terazosin; Propafenone; Apixaban; Ticagrelor; Azilsartan medoxomil; Pitavastatin; Aliskiren;</td>
<td>0.435</td>
</tr>
<tr>
<td>Bendroflumethiazide</td>
<td>8</td>
<td>Niacin; Clofibrate; Simvastatin; Fluvastatin; Rosuvastatin; Pitavastatin;</td>
<td>0.29</td>
</tr>
<tr>
<td>Benazepril</td>
<td>7</td>
<td>Niacin; Clofibrate; Acetazolamide; Donepezil; Fluvastatin; Gemfibrozil; Fenofibric acid;</td>
<td>0.25</td>
</tr>
<tr>
<td>Disopyramide</td>
<td>7</td>
<td>Trandolapril; Clofibrate; Simvastatin; Telmisartan; Irbesartan; Atorvastatin; Aliskiren;</td>
<td>0.25</td>
</tr>
<tr>
<td>Rosuvastatin</td>
<td>6</td>
<td>Papaverine; Terazosin; Propafenone; Azilsartan medoxomil; Aliskiren; Enalaprilat;</td>
<td>0.21</td>
</tr>
<tr>
<td>Moexipril</td>
<td>6</td>
<td>Acetazolamide; Donepezil; Fenofibrate; Fluvastatin; Gemfibrozil; Ticagrelor;</td>
<td>0.21</td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>6</td>
<td>Irbesartan; Fluvastatin; Captopril; Hydralazine; Pitavastatin; Aliskiren;</td>
<td>0.21</td>
</tr>
<tr>
<td>Clopidogrel</td>
<td>6</td>
<td>Perindopril; Acetazolamide; Methyldopa; Terazosin; Aliskiren; Fenofibric acid;</td>
<td>0.21</td>
</tr>
<tr>
<td>Valsartan</td>
<td>5</td>
<td>Ibutilide; Amlodipine; Niacin; Clofibrate; Aliskiren;</td>
<td>0.18</td>
</tr>
<tr>
<td>Ibutilide</td>
<td>5</td>
<td>Niacin; Guanabenz; Fluvastatin; Terazosin; Pitavastatin;</td>
<td>0.18</td>
</tr>
</tbody>
</table>
5. EXPERIMENTS AND RESULTS

<table>
<thead>
<tr>
<th>Drug</th>
<th>Count</th>
<th>Associated Drugs</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donepezil</td>
<td>5</td>
<td>Dopamine; Terazosin; Hydralazine; Aliskiren; Fenofibric acid;</td>
<td>0.18</td>
</tr>
<tr>
<td>Guanabenz</td>
<td>4</td>
<td>Clofibrate; Donepezil; Fluvastatin; Prasugrel;</td>
<td>0.14</td>
</tr>
<tr>
<td>Milrinone</td>
<td>4</td>
<td>Niacin; Fluvastatin; Propafenone; Ticagrelor;</td>
<td>0.14</td>
</tr>
<tr>
<td>Ticagrelor</td>
<td>4</td>
<td>Aliskiren; Sacubitril; Enalaprilat; Fenofibric acid;</td>
<td>0.14</td>
</tr>
<tr>
<td>Papaverine</td>
<td>3</td>
<td>Prasugrel; Ticagrelor; Pitavastatin;</td>
<td>0.10</td>
</tr>
<tr>
<td>Fenofibrate</td>
<td>3</td>
<td>Papaverine; Terazosin; Azilsartan medoxomil;</td>
<td>0.10</td>
</tr>
<tr>
<td>Aliskiren</td>
<td>3</td>
<td>Vorapaxar; Edoxaban; Fenofibric acid;</td>
<td>0.10</td>
</tr>
<tr>
<td>Fluvastatin</td>
<td>3</td>
<td>Papaverine; Azilsartan medoxomil; Aliskiren;</td>
<td>0.10</td>
</tr>
<tr>
<td>Gemfibrozil</td>
<td>3</td>
<td>Ticagrelor; Azilsartan medoxomil; Aliskiren;</td>
<td>0.10</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>3</td>
<td>Dopamine; Sitagliptin; Aliskiren;</td>
<td>0.10</td>
</tr>
<tr>
<td>Pitavastatin</td>
<td>3</td>
<td>Aliskiren; Vorapaxar; Enalaprilat;</td>
<td>0.10</td>
</tr>
<tr>
<td>Dipyridamole</td>
<td>2</td>
<td>Dopamine; Fenofibric acid;</td>
<td>0.072</td>
</tr>
<tr>
<td>Trandolapril</td>
<td>2</td>
<td>Simvastatin; Fluvastatin;</td>
<td>0.072</td>
</tr>
<tr>
<td>Eprosartan</td>
<td>2</td>
<td>Prasugrel; Ticagrelor;</td>
<td>0.072</td>
</tr>
<tr>
<td>Lisinopril</td>
<td>2</td>
<td>Fluvastatin; Ticagrelor;</td>
<td>0.072</td>
</tr>
<tr>
<td>Dronedarone</td>
<td>2</td>
<td>Azilsartan medoxomil; Fenofibric acid;</td>
<td>0.072</td>
</tr>
</tbody>
</table>
### 5. EXPERIMENTS AND RESULTS

<table>
<thead>
<tr>
<th>Medication</th>
<th>Count</th>
<th>Associated Medications</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pindolol</td>
<td>2</td>
<td>Rosuvastatin; Ticagrelor;</td>
<td>0.072</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>2</td>
<td>Pitavastatin; Aliskiren;</td>
<td>0.072</td>
</tr>
<tr>
<td>Captopril</td>
<td>2</td>
<td>Prasugrel; Ticagrelor;</td>
<td>0.072</td>
</tr>
<tr>
<td>Guanfacine</td>
<td>2</td>
<td>Fluvastatin; Ticagrelor;</td>
<td>0.072</td>
</tr>
<tr>
<td>Apixaban</td>
<td>2</td>
<td>Aliskiren; Enalaprilat;</td>
<td>0.072</td>
</tr>
<tr>
<td>Hydralazine</td>
<td>1</td>
<td>Ticagrelor;</td>
<td>0.036</td>
</tr>
<tr>
<td>Triamterene</td>
<td>1</td>
<td>Niacin;</td>
<td>0.036</td>
</tr>
<tr>
<td>Amlodipine</td>
<td>1</td>
<td>Niacin;</td>
<td>0.036</td>
</tr>
<tr>
<td>Labetalol</td>
<td>1</td>
<td>Rosuvastatin;</td>
<td>0.036</td>
</tr>
<tr>
<td>Sotalol</td>
<td>1</td>
<td>Fenofibrate;</td>
<td>0.036</td>
</tr>
<tr>
<td>Dopamine</td>
<td>1</td>
<td>Atorvastatin;</td>
<td>0.036</td>
</tr>
<tr>
<td>Sitagliptin</td>
<td>1</td>
<td>Aliskiren;</td>
<td>0.036</td>
</tr>
<tr>
<td>Fosinopril</td>
<td>1</td>
<td>Ticagrelor;</td>
<td>0.036</td>
</tr>
<tr>
<td>Esmolol</td>
<td>1</td>
<td>Gemfibrozil;</td>
<td>0.036</td>
</tr>
<tr>
<td>Azilsartan medo-</td>
<td>1</td>
<td>Pitavastatin;</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>doxomil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fenofibric acid</td>
<td>1</td>
<td>Candesartan;</td>
<td>0.036</td>
</tr>
<tr>
<td>Ramipril</td>
<td>1</td>
<td>Fluvastatin;</td>
<td>0.036</td>
</tr>
<tr>
<td>Isradipine</td>
<td>1</td>
<td>Fluvastatin;</td>
<td>0.036</td>
</tr>
<tr>
<td>Perindopril</td>
<td>1</td>
<td>Ticagrelor;</td>
<td>0.036</td>
</tr>
<tr>
<td>Diltiazem</td>
<td>1</td>
<td>Acetazolamide;</td>
<td>0.036</td>
</tr>
</tbody>
</table>
Fig. 5.4.1: PS3N network diagram for the cardiovascular drugs. In this network, there are only 87 cardiovascular drugs are available among the 489 drugs we considered for the network diagram. This network is representing the overall interactions based on PS3N network. The red lines are indicating the new interactions.

5.4.1.1 PS3N DDI Network for cardiovascular drugs

To visualize the performance of the PS3N network on cardiovascular drugs, we considered a subset of protein sequence and protein structure networks into consideration to predict their labels. After calculating the network information like the predicted labels, cardiovascular interactions, and non-cardiovascular interactions, we can see the overall interactions in 5.4.1. the gray regions are representing all possible DDI's available or known to us. The red dotted lines are showing newly predicted interactions among different drugs.

As the visualization for this big graph is difficult to find out specific information for individual drugs, we showed an example in 5.4.2 to get some idea about how we can utilize our model to find out the specific queries regarding drug interactions. In this example, we can see that Alteplase (popularly used for cardiovascular disease to dissolve blood clots) is interacting with drugs like hydrochlorothiazide which normally increase blood pressure and that is not good for cardiovascular patients.
5. EXPERIMENTS AND RESULTS

Fig. 5.4.2: This diagram is representing the network for drug Alteplase. The green node representing Alteplase and the red dotted lines are representing the new inter-cardiovascular drug interactions. Among the 87 cardiovascular drugs, there are new 6 interactions for Alteplase.

5.4.1.2 PSSCSSN DDI Network for cardiovascular drugs

In the PSSCSSN model, there are three different types of information were utilized to create the similarity network. So when we check the model performance in visualization, there are lots of new interactions found. As we got some statistical idea about this on Table. 5.4.2. Here we will explore the newly discovered interactions in Fig. 5.4.3, which is representing a circular interaction diagram. the center of the diagram represents the drug Pravastatin and all other nodes have edges from the center. we can see 176 different edges showing newly discovered interactions with the drug.

In Fig. 5.4.4, there are information’s about inter-cardiovascular drugs for Pravastatin. There are different edge attributes are available for this as well such as numerical values of similarities which could potentially be used for identify different aspects of DDI’s like severity, causality, etc.
5. EXPERIMENTS AND RESULTS

Fig. 5.4.3: PSSCSSN Network diagram for a particular cardiovascular drug Pravastatin. The network contains an overall 722 drugs. It is very difficult to show every interactions from all the nodes. So we considered a subset of interactions which are newly found and interacting with Pravastatin. Green lines represent the inter-cardiovascular interactions with Pravastatin.

Fig. 5.4.4: Cardiovascular interaction diagram for Pravastatin. All other drugs interactions are discarded for better visualization and understanding.
CHAPTER 6

Discussion

The main objective of this work is to propose a new computational model for DDI prediction utilizing the genetic information about drug protein targets. Our work has given a promising direction for addressing DDI prediction problems. We showed different ways of creating the feature space to identify the interaction between a pair of drugs. Roughly, we identified drugs with information on protein structures, and drugs with information on the protein sequence. We created the labeled feature space by utilizing the interaction information available in DrugBank. The combination of the structure and sequence information resulted in 904 drugs. Unlike previous methodologies, we considered only protein sequence and structure similarity networks for the first time to predict drug interactions. In addition, our similarity network computation technique allows extracting important protein features in terms of different distance measures.

Our proposed model for cardiovascular drug interactions is likely the first to take into account all sorts of interactions in cardiovascular issues. Most of them worked on patent information, utilizing multiple-coexisting patient circumstances on cardiovascular disease, according to our findings in a literature analysis on cardiovascular drug interactions. Those studies largely looked at common drugs to discover their clinical importance and explanation. In terms of similarity values, our suggested method considers 722 pharmaceuticals with a knowledge base of protein sequence, protein structure, and chemical structure. We were able to generate network data from the prediction as well. This vast amount of information could be crucial for future cardiovascular disease studies.
6. DISCUSSION

The major drawback of our work is the lack of availability of drug information. Due to the different sources, it’s difficult to get all types of information for the same drug. As we mainly focused on Drugbank and Protein Data Bank (PDB) with a focus on SIDER and HIPPIE for Side-effects and PPI, it was a challenge to find the commonality between the different datasets. Moreover, the datasets have significantly more unknown interactions than known interactions. Thus, this creates a problem of data imbalance, especially if we do not consider appropriate thresholds for the unknowns. However, the time and space complexity for feature space generation is significant and that will need to be addressed in the future.
CHAPTER 7

Conclusion

In this study, we proposed a novel drug-drug interaction detection mechanism. The proposed model is divided into three major chunks. The first is focused on building the similarity profiles from Drug Bank and PDB. The second is the creation of an integrated similarity network (PS3N) about drugs, using information about their protein targets, namely, the protein sequences and protein structures of such targets. The third component is how information from the integrated network is used to develop a deep neural network model for improved prediction of the potential drug interactions. We compare the results produced using the proposed PS3N in a deep learning framework with results from other recent machine learning-based approaches. The comparisons showed that our proposed methodology is quite competitive with respect to the state-of-the-art, at times outperforming the state-of-the-art methods. Though the computational complexity is high for the pre-processing, there are still opportunities to improve the performance of the model and also improve the datasets as well. In our proposed methodology, we showed a new approach to dealing with the DDI prediction problem, by exploiting genetic information about the drug protein targets, in particular, information about their protein sequence and protein structure.

After that, we created the PSSCSSN model for cardiovascular medications in order to find previously unknown drug interactions. The PSSCSSN model was created using protein sequence, protein structure, and chemical structure drug similarity networks. We were able to find a considerable number of previously unreported interactions with cardiovascular medications using our proposed approach. Interactions with non-cardiovascular medicines were also detected using the model.
The PSSCSSN model has generated a knowledge base for cardiovascular drug interactions, which could aid cardiovascular disease research in a variety of ways, including clinical decision making, pharmacist training, and electronic prescription to reduce patent risk. Future research could look into how the general approach could be applied to other adverse drug effects besides DDIs. Also, with a complete understanding of various drug-drug interactions, the work might be expanded to assess the frequency, level, and risk factors of cardiovascular disease patients. Using the PS3N and PSSCSSN models, we hope to locate COVID-19 and learn more about the negative effects of immunizations and other medications.
REFERENCES


[8] Andrej K., Polonca F., and Brane L. (2018). Predicting potential drug-drug in-


56


59