APPLICATION OF MACHINE LEARNING

IN DRUG DISCOVERY

by

Susmitha Sri Kadiyala

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Mehrdad Nourani, Chair

Dr. Lakshman S. Tamil

Dr. Jorge A. Cobb

Dedicated to my Parents, Teachers and my Brother.

Thank you for always being there for me.

APPLICATION OF MACHINE LEARNING IN DRUG DISCOVERY

by

SUSMITHA SRI KADIYALA, B.Tech.

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN

COMPUTER ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

December 2018

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Professor Mehrdad Nourani for his excellent guidance, caring and for providing me an outstanding atmosphere throughout my research. I am very grateful for the opportunity to work in his research lab.

I would like to thank my committee members, Professor Lakshman S. Tamil and Professor Jorge A. Cobb for serving in my M.S. committee and for letting my defense be an enjoyable moment and for their brilliant comments and suggestions.

I also want to thank Hafez Eslami Manoochehri and Javad Birjandtalab, Dr. Nourani's PhD students for their encouragement and support.

I would also like to thank my parents for their love, kindness and support. They were always supporting me and encouraging me with their best wishes.

November 2018

APPLICATION OF MACHINE LEARNING

IN DRUG DISCOVERY

Susmitha Sri Kadiyala, MS
The University of Texas at Dallas, 2018

Supervising Professor: Dr. Mehrdad Nourani, Chair

Drug Discovery is a highly complicated process. On average, it takes 6 to 12 years to manufacture a drug and have the product released in the market. Even after a huge investment of money, time and hard work, one cannot assure the success of the drug after its release. The recent advancement in the field of machine learning helps us to reduce the risk in this field of science. This thesis aims at analyzing the applications of machine learning in the field of bio-medical science. Usage of a simpler organism for the implementation of the experiments is highly convenient. Therefore, a machine learning model to predict the chemical compounds effect on aging of Caenorhabditis elegans was proposed using the Drug Age database. This database includes the features of Molecular Descriptors and Gene Ontology.

In this work, a new feature selection scheme is proposed for an efficient classification task using random forests. We explain the benefits of our feature selection method in comparison with the base-line support vector machine and artificial neural network classifiers. Secondly, another application of machine learning which is presented in the work is the prediction of Drug-Target Interaction using Weisfeiler-Lehman Neural Machine. Prediction of a possible interaction between a drug and a target enables the biochemists to speed up the process of target validation and discovery. A public-domain data set which corresponds to four different

target protein types is used for the analysis purpose. The algorithm aims at creating a subgraph from the network formed by the drugs and targets which is then taken through graph labeling, resulting in the formation of an adjacency matrix. This matrix defines the presence of an interaction used for training a model. The results of the proposed method out performed the standard state of art approaches like the similarity based methods in terms of AUC.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Drug Discovery Process

Drug discovery is a very complicated process as it involves a huge investment of time and money. On average, it takes 6 to 12 years with an investment of 500 million to 1 billion (in US dollars) to identify a drug for fighting against a target. However, even after a huge struggle, the success rate is very low. Many long-term research projects may end up fruitless resulting in wastage of enormous efforts. Blockbuster drugs are the drugs that are prescribed for the common medical problems like cold, diabetes, high blood pressure, asthma and flu. They are extremely profitable in the pharmaceutical industry. They bring revenues greater than 1 billion per year and a profit of more than 1 million a day (in dollars). However, it can also result in problems for the company if the drug shows any side effects. Usually, the patents on drugs expire resulting in competition from less expensive equivalents. The process of drug discovery is therefore highly complicated and risky activity but is always motivated by the benefits it could do to millions of people suffering from various diseases. The detailed process of drug discovery, illustrated, in Figure 1.1 is as follows:

1. The first stage is the target discovery. In this stage, we decide on the target on which the drug in development should act upon to suppress the growth of the disease. The target gives us the deeper understanding of the genes, proteins, Ribonucleic Acid (RNA) or anything that get effected when attacked by a parasite.

2. The second stage is the target validation phase. In this phase, the discovered target is validated to make sure that the drug in development deals with the correct target.

3. The third stage is the lead discovery. This phase involves in synthesizing and isolating the designed chemical compounds meant to interact with the specified targets. This stage involves the use of chemistry, assay development for screening the selected compounds.

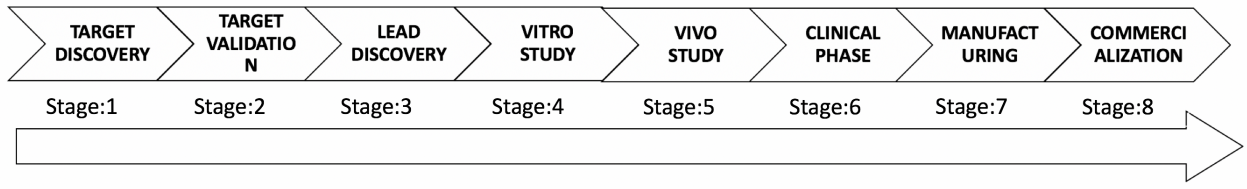

Figure 1.1. Process of Drug Discovery

4. The fourth stage is the vitro study phase where testing of the drug identified is done prior to testing on animals. This study analyzes the affinity of the drug towards the target. The mechanism in which the drug acts is studied.

5. The fifth stage is the vivo study in which testing on animals is involved. Frequently used animals are rats or mice. This allows the chemists to understand the working of the drug in a biological model. It provides an understanding of drug metabolism, clearance, immune response, etc. giving a detailed information back to the chemist. It provides a deeper understanding of the behavior and functional imaging of the response of the disease to the drug developed.

6. The sixth stage is the clinical trial phase where the drug is tested on humans. If the drug shows efficiency and meets the required purpose, it proceeds with the last step.

7. The seventh stage is the manufacturing stage. Once the drug successfully clears all the tests on it, the chemists then start manufacturing it for usage by people.

8. The final stage in the process is commercialization. Once the drug passes the review, and approved by Food and Drug Administration (FDA), it can be prescribed by physicians clinically. This drug is then made commercially available for people to purchase.

The successful completion of all these phases takes many years. Research is going on to improve the speed in this process with higher efficiency to fight against a disease by the drug.

1.2 Applications of Machine Learning

Machine Learning is a field of computer science which focuses on creating computational algorithms that can learn complex patterns for a given set of inputs with their corresponding labels and predict and classify the new samples into any of the existing labels. The input data is referred to as training data and the new sample set is the test data. We use this data to train a model for classifying the new data into a predefined class label. The ultimate goal of the various computational methods is to classify and to predict the behavior of test data using the train data with higher accuracy.

There exists a procedural similarity with data mining and predictive modeling in machine learning. Both data mining and predictive modeling require a search through data to identify various possible patterns for modeling. The usage of machine learning by the shopping related websites, where there will be suggestions related to the activity of the customers, is an application of machine learning in the daily life. Other applications of machine learning include spam filtering, social networking, online advertising, etc. (Guzella and Caminhas, 2009) (Benchettara et al., 2010) (Goodfellow et al., 2016).

Machine learning algorithms can be classified as supervised, unsupervised, semi-supervised or reinforcement learning (Pedregosa et al., 2011). Supervised learning requires the provision of input data as well as the desired output. It also takes care of providing the accuracy predictions during training phase. The features, instances and the model to be used are to be determined prior to application of the algorithm on new test data. On achieving the acceptable performance level, the learning can be halted. The supervised learning problems can be grouped as either classification or regression problems. A classification problem is the one in which the output is a category. For example, YES or NO. The regression problem has a real valued output like weight, length, etc. On the other hand, unsupervised algorithms do not require to be trained with desired outcome. They use an iterative approach to model the underlying distribution providing a scope to learn more about the data.

These problems are grouped as clustering or association problems. In clustering, we define the inherent groups in the data and in association we try to define the rules for understanding the large data (Domingos, 2012). There is also semi-supervised learning in which we have an input data and a part of it is labeled. Many of the real-world problems fall into this category. To solve these problems, both supervised and unsupervised learning methods are employed. Reinforcement learning uses the observations gathered from environmental interaction. Reinforcement learning algorithm learns from the environmental setup iteratively until it is sure of a reduced risk. It uses a feedback signal called reinforcement signal to learn the behavior of environment.

There are various machine learning algorithms (Learning, 2012), some of the popular approaches are,

- Decision Trees: It uses a tree like graph model consisting of observations about certain decisions and their possible outcomes. Pruning can help improve the performance of a tree by removing the branches with low importance. This reduces the complexity of the tree as well as the over-fitting.

- Naive Bayes Classification: These classifiers are based on the Bayes theorem and are particularly used when the dimensionality of inputs is high. It is the simplest model outperforming many complicated classification methods.

- K-means Clustering: This algorithm helps in grouping the data, with the group number denoted by K. It assigns each data point to a group by iteration based on features provided. These data points are then clustered based on the similarity of the feature. The results of K-means clustering are the centroids of the K clusters and labels for the data.

- Logistic Regression: It is a statistical analysis method for analyzing the data set with one or more than one independent variable helping in determining an outcome. Logistic

regression is a predictive analysis used to describe data for explaining the relation between a binary variable and other independent variables.

- Support Vector Machines: It classifies the training data by trying to model it into a decision boundary while maximizing the margin between classes. It uses kernel function when there is no possibility for linear separation of data.

- Neural Networks: It is a parameterized non-linear algorithm consisting of multi-layer perceptrons at each layer for classification of input data. The numeric value of perceptrons and hidden layers in the model decide the accuracy.

1.3 Machine Learning in Drug Discovery

Machine learning methods have been used in the area of drug discovery since 1962. Machine learning methods help in understanding the complex biological systems by enabling us to capture all the relevant features. Usage of various prediction models to improve the speed in process of drug discovery is recently extensive in this area of research. The algorithms, used by various computational methods, enable us to reveal answers to questions that pose a greater challenge to the chemists. They help the chemists to analyze, predict and model many biological responses for an accurate drug design. The machine learning algorithms learn complicated patterns with the help of the annotated data to predict the annotations of new test data samples (Burbidge et al., 2001). Machine learning is used in protein structure prediction, protein function prediction, genome association and so on. It helps in understanding the properties of various compounds like solubility, binding and assays related to targets, etc.

Despite the success, implementing machine learning in the area of drug discovery is never an easy task. Contrary to the other fields, drug discovery poses many challenges related to identifying the right representation for the subjects in a drug, like the molecules and their

complexes which play a major role in achieving the goal. One of the major challenges is the shortage in the description of the bio-activity. The way to use the available data for reaching the goal is highly important.

So, figuring out the correct representation is always the most challenging. The machine learning methods are data dependent; especially the training data. This is even more challenging as the data obtained for most of the predictions has high noise levels, uncertainty and most of the times they are inconsistent. The chemical experiments performed may result in a data that is sparse and unbalanced making it even more difficult. In the recent times, computational methods to address these difficulties are being developed. There are many possibilities for expanding the help of machine learning in understanding the bio-activity data leading to an acceleration in the process of drug discovery and development

1.4 Contribution and Thesis Organization

This work mainly focuses on the target discovery stage of the drug discovery process. Chapter 2 of the thesis is about an efficient feature selection method for the determination of chemical compounds effect on aging of C.elegans, an organism that is used for testing purposes by biochemists. Selection of most important features among Molecular Descriptors and Gene Ontology is done using Particle Swarm Optimization and Correlation Based Feature Selection method respectively. The results obtained through the proposed method are better than the previous techniques.

Chapter 3 focuses on the methodology for prediction of a possible interaction between a drug and target. In this work, we use the data of the four different target protein types. We use Weisfeiler-Lehman Neural Machine algorithm for DTI prediction. The work is mainly on the selection of the negative unknown samples of the DTI. We use two different sampling techniques for this purpose. Finally, in Chapter 4, we outline the conclusion of the work and its future prospective.

CHAPTER 2

FEATURE SELECTION TO PREDICT COMPOUND'S EFFECT ON AGING

Acknowledgement: The main part of this chapter has been reported in this paper: Eslami Manoochehri Hafez, Susmitha Sri Kadiyala, Javad Birjandtalab, and Mehrdad Nourani. "Feature Selection to Predict Compound's Effect on Aging", In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 419-427. ACM, 2018.

2.1 Overview

The process of aging results in many age-related disorders. Investigating the changes due to aging at cellular level and understanding the chemical impacts among the anti-aging compounds is of high interest in drug discovery and in personalized drugs research. In this work, a model to predict the effect of chemical compounds on lifespan of Caenorhabditis elegans is proposed. The data from DrugAge database was used for analysis (Eslami Manoochehri et al., 2018). This data includes chemical compounds that affect the lifespan of model organisms by using the chemical descriptors and gene ontology as features. We used a new feature selection scheme based on particle swarm optimization and correlation-based feature selection for selecting the most relevant features for classification. This method achieved higher performance compared to the existing methodologies. The advantages of the proposed feature selection method over the other methodologies are discussed and the results obtained by random forest with base-line support vector machine artificial neural network classifiers are compared with the previous methodologies.

A wide range of molecular and cellular changes over time results in the biological process of aging. The declination in biological function due to aging leads to gradual decrease in physical and mental capacity. This could increase the risk of diseases like cancer and other

neuro-generative diseases. These biological changes are not monotonic, so understanding their impacts can be very challenging. exploring the aging processes and understanding them would enable us to promote healthy aging preventing many age-related diseases. Anti-aging drug discovery is a biggest challenge that can help us to fight against the advancement of diseases. Intra-cellular signaling pathways modulated by dietary, nutraceutical, genetic and pharmacological interventions can be considered to slow down the aging process.

2.1.1 Prior Works

Aging studies are made on organisms such as Caenorhabditis elegans (C. elegans) worm, yeast, mice and fly (Buffenstein et al., 2008). Among them, C. elegans is highly used for conducting assessments on anti-aging and drug intervention due to its short lifespan, stereotypical development and small size (Kaletta and Hengartner, 2006) (Lucanic et al., 2013) (Carretero et al., 2017). There has been a decent increase in the number of works done on the identification of compounds that increase the lifespan. Latest research enabled us to develop compounds that enhance the longevity of caloric restriction (Fontana et al., 2010). In (Calvert et al., 2016), a pharmacological network is constructed by Ye et al. with the help of a connectivity map to identify drugs with overlapping gene expression profiles. It helped to identify 60 compounds that improve longevity for C. elegans (Ye et al., 2014). Ranking of drug like compounds to modulate aging in C. elegans has been proposed in (Ziehm et al., 2017). This ranking method is based on genetic information, information on proteins associated with aging in an organism, 3D protein structure, homo-logy and sequence conversation between them and compound activity information. In (Ding et al., 2017) and (Carretero et al., 2015), a review on anti-aging and pharmacological compound classification on C. elegans lifespan has been discussed.

The process of drug discovery includes several steps from selecting chemical compound candidates to clearing all drug requirement tests. It is also time consuming, expensive and

labor-intensive. Recent advancements in machine learning is helping out in prediction of the chemical properties in drug discovery community. In (Zernov et al., 2003), Support Vector Machine is used for predicting the drug-likeness and agrochemical-likeness for large number of compounds. In (Putin et al., 2016), human chronological age was predicted using Deep Neural Network. In (Fabris et al., 2018), prediction of aging related genes was done using random forests.

Different possibilities of involvement of machine learning and artificial intelligence in longevity and anti-aging discoveries was discussed in (Batin et al., 2017). Machine learning was not greatly used for compound prediction on longevity. To the best of our knowledge, only a few studies tried to identify chemical compounds that effects longevity (Putin et al., 2016)(Barardo et al., 2017) (Fabris et al., 2017). Barardo et al. worked on classification of chemical compounds into two classes which are the ones that effect longevity and the ones that do not, using random forests (Barardo et al., 2017). They used random forest feature importance as a parameter to select the most important features among the two types of features which are chemical descriptors, gene ontology terms. Then, they implemented random forests for classification. Their results showed that impact of chemical descriptors was high compared to GO. However, using both chemical descriptors and Gene Ontology slightly improves the classification accuracy.

2.1.2 Feature Selection

Information provided by features in a data set play a major role in classifying each instance into different classes. In many cases, a compromise on relevant and redundant features is observed in the data. Redundant features slow down the process of learning resulting in a decrease in accuracy. Feature selection helps us to identify and eliminate these redundant features helping us to improve the performance of the classification (Dash and Liu, 1997) (Xue et al., 2013). Exponential increase in the size of search space with the size of features

makes feature selection much more challenging. Due to this, search methods such as greedy search or heuristic-based search have been proposed (Mao and Tsang, 2013). But these methods are prone to suffer from the local optima problem. Feature selection can be done using two approaches: 1) methods that are independent of a learning algorithm and do not consider classifier performance (Moradi and Rostami, 2015), and 2) wrapper methods which learns an algorithm in the feature selection process (Gasca et al., 2006) (Wang et al., 2008). Wrapper methods can enable us to achieve higher predictive accuracy (Dash and Liu, 1997).

2.1.3 Key Contributions

This work focuses on analyzing the effect of chemical compound on lifespan of C. elegans. DrugAge data set (Barardo et al., 2017) has molecular descriptors and gene ontology terms as features for building a machine learning model that enables us to classify a new chemical compound based on its capability to affect the lifespan. Two different feature selection methods, a swarm-based wrapper method for molecular descriptors and a correlation-based feature selection method for Gene Ontology (GO) features were used. Feature selection is highly important, and it mainly serves two objectives 1) maximizing the classification performance and 2) minimizing the number of features. To satisfy the above two objectives, the feature selection methods were modified to select a proper feature subset. We applied Random Forest classifier on the selected features and a good prediction performance is measured using Area Under the Curve (AUC) score and accuracy was achieved. This work was compared with the existing work in (Barardo et al., 2017) and it showed a better performance with similar number of MD features and lesser GO features. The results were also compared with two other classification methods, Artificial Neural Network and Support Vector Machine.

The organization of the chapter is as follows. We present our main methodology in Section 2.2 which includes pre-processing, feature selection and classification steps. Then, in Section 2.3, we present the experimental results of our approach.

2.2 Methodology

The aim of the work is to identify the effect of chemical compounds on longevity of C. elegans. The two features used in this work are evaluated using different measures. The Molecular Descriptors are calculated from the chemical structure enabling us to define the relation between chemical structure of the compound and its biological properties. There are a total of 268 MD's and they are real values. The GO features on the other hand are derived from the proteins which interact and get targeted by each of the compounds. The Feature size of GO is 13,338 and they take binary values of 0 or 1 representing the presence of an annotation of the instance with the corresponding GO term. (see Subsection 2.3.1 for more information about the data).

Figure 2.1, shows the methodology followed for the work. In the work, the raw data is first normalized followed by the feature selection which is done for the selection of appropriate features to feed the classifier. Two different feature selection methods are proposed for the two types of features. A wrapper method, modified Particle Swarm Optimization (PSO) is used for subset selection of Molecular Descriptors (MD) features. Since this method is expensive computationally, it was not applied on GO features as they are very high in number compared to the MD's. To work with the large size of GO features, a filter method Correlation-Based Feature Selection (CFS) is used. The advantage of filter methods is that they are independent of the classifier performance. Once the features are selected, they are fed to a random forest classification block which helps us to build a predictive model.

2.2.1 Feature Selection for Molecular Descriptors

Initially, we normalize the data of the MDs re-scaling it to a value between $[0, 1]$. This eliminates the scaling effect. We normalize only the MDs and not the GOs as they already take binary values 0 or 1.

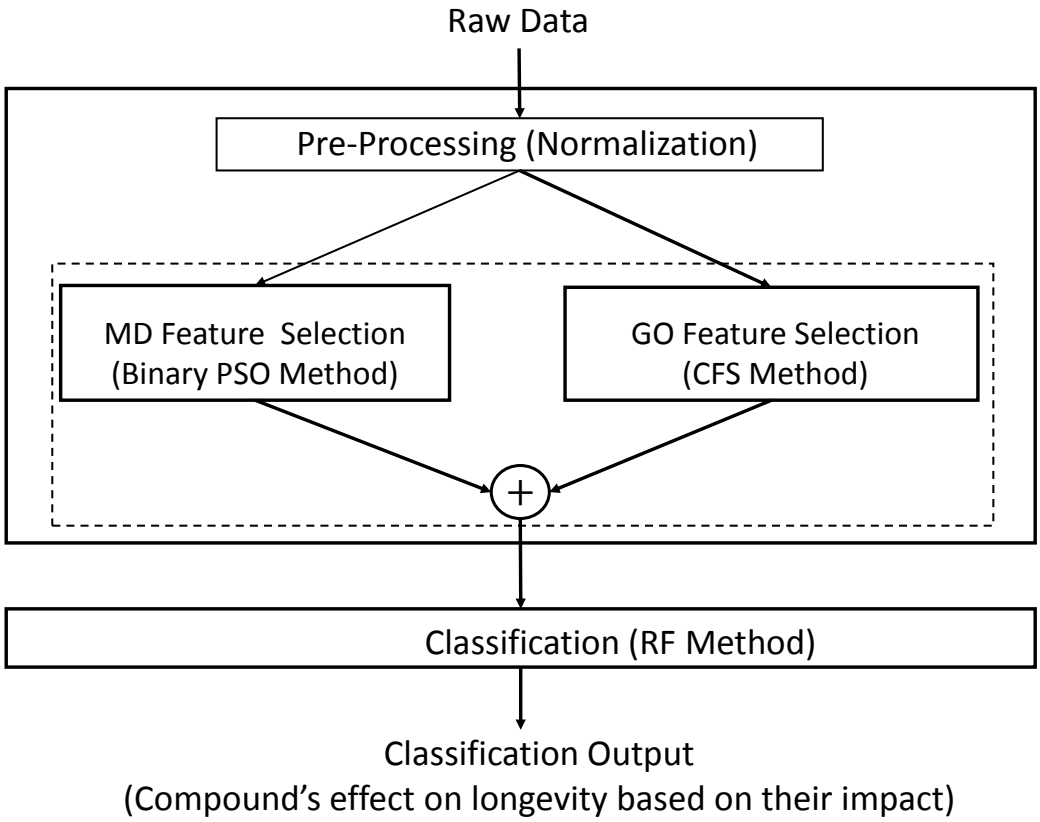

Figure 2.1. Schematic Model

Exhaustive search is not suggested as the selection of most appropriate feature subset becomes challenging with increase in size of the feature space. Thus, heuristic methods or optimization methods are better for the feature selection. Despite the higher performance compared to filter methods, wrapper methods suffer from the drawback of high computational costs. Swarm based Evolutionary Computation (EC) techniques like Genetic Algorithm (GA) (Siedlecki and Sklansky, 1993), Ant Colony Optimization (ACO) (Tabakhi et al., 2014) and PSO (Moradi and Gholampour, 2016) are known for their global search potential. PSO among them is computationally economic (Xue et al., 2016).

Particle Swarm Optimization(PSO)

(Kennedy, 2011) (Shi and Eberhart, 1998) is a less expensive recent EC technique, it can converge faster than some other EC algorithms such as GA. Thus, PSO is being used as an effective technique for the process of feature selection, specially when the number of features are large (Xue et al., 2013). In addition to PSO's convergence speed, it offers better results compared with those of the other stochastic optimization methods (Kennedy, 2011)(Xue et al., 2016).

Representing each candidate as a particle in swarm is the basic principle of PSO method. Each particle i has a position in the search space, which is represented by a vector $\mathbf{x}_i = (x_i^1, x_i^2, ..., x_i^D)$, where D denotes the dimensionality of the search space. These particles move in the search space in search of the optimal solutions. The velocity of each particle i is represented by $\mathbf{v}_i = (v_i^1, v_i^2, ..., v_i^D)$. During the movement, each particle updates its position and velocity according to its own and neighbors' experiences. The best of the particle i's previous state is given a the personal best x_i^{best}, and the best position based on the population is denoted by x_g^{best}. Based on x_i^{best} and x_g^{best}, PSO works by searching a valid solution for the position and velocity for each particle by following the equations given below:

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \tag{2.1}$$

$$v_i^d(t+1) = wv_i^d(t) + c_1 r_{1i}(x_i^{best} - x_i^d(t)) + c_2 r_{2i}(x_g^{best} - x_i^d(t)) \tag{2.2}$$

where the tth iteration in the process of evolution is denoted by t. $d \in D$ is the dth dimension in the given search space and i denotes the ith particle. w is inertia coefficient used to control the effect of earlier velocities on the current one. If $w < 1$, we give more importance to exploitation over exploration and then to the positions. Conversely, if $w > 1$, we give more importance to current best positions of particles. c_1 and c_2 are acceleration coefficients (learning factors) where c_1 implies the degree of self-confidence of particle while c_2 implies

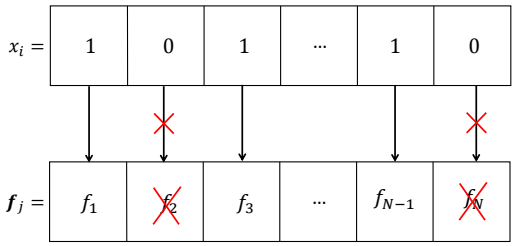

Figure 2.2. Feature encoding.

the degree of confidence in candidate solution of the swarm. r_1 and r_2 are random values uniformly distributed in [0, 1]. Note that x_i^{best} and x_g^{best} are local best position for x_i and global best positions, respectively, for dth dimension up to iteration t. Usually to keep the coherence of the swarm, the velocity is kept within limits by a predefined maximum velocity, v_{max}, so $v_i^d(t+1) \in [-v_{max}, v_{max}]$. Typically, in PSO a fitness function F (in case of feature selection, classification error or performance) is defined. The algorithm stops when a predefined criterion is met, which could be an appropriate fitness value or a predefined maximum number of iterations.

Binary PSO Feature Selection

A special case of PSO (i.e Binary PSO) can be applied for feature subset selection. Let's assume $\mathbf{x}$ is a vector of binary variable so $\mathbf{x} \in \{0, 1\}$. The length of vector $\mathbf{x}$ must be equal to the size of feature space. Now an encoding scheme can be done by corresponding each variable $x \in \mathbf{x}$ to a feature $f \in \mathbf{f}$. Thus, if a feature f_i is to be selected then $x_i = 1$, otherwise $x_i = 0$. This encoding is shown in Figure 2.2. The binary PSO feature selection method keeps executing the following steps until it meets the stopping criteria, either convergence or reaching to a predefined iteration limit:

- Initialization: In this step, parameters such as number of iterations and number of particles are set. Additionally, each particle's position and velocity vectors as well as parameters in Eqn. 2.2 are initialized.

- Particle evaluation: In this step, each particle which represents a potential feature subset, is evaluated by a fitness function. In the basic PSO, a single objective function is used to minimize the classification error. This function is usually provided by a supervised classification method and does not take into account minimizing the number of features. Elimination of redundant features can help attaining the same performance when smaller number of features are considered. However, reducing the number of features and having higher performance are conflicting. We define the objective function based on (Vieira et al., 2013) as follows:

$$F(X) = \alpha(1 - P) + (1 - \alpha)(1 - \frac{N_f}{N_t}) \tag{2.3}$$

where P is the classifier performance (see Section 2.2.3) on the selected feature subset and N_f and N_t are the size of tested feature subset and total feature size, respectively. Various measures have been designed to evaluate the classifier performance such as accuracy, precision, Hamming loss and etc. (Zhang and Zhou, 2014). To deal with the skewed sample distribution of the data toward negative class, instead of accuracy which typically is used as a classifier performance metric, we chose Area Under the Curve (AUC) score for P (see section 2.3.2). $\alpha \in [0, 1]$ is a hyper-parameter to control the trade-off between performance and size of feature subset.

- Computing local and global best: After computing the fitness value for each particle, at this step, $\mathbf{x}_i^{best}$ and $\mathbf{x}_g^{best}$ are updated as follows:

$$\mathbf{x}_i^{best} \longleftarrow \mathbf{x}_i, \quad if\ F(\mathbf{x}_i) > F(\mathbf{x}_i^{best}) \tag{2.4}$$

$$\mathbf{x}_g^{best} \longleftarrow \mathbf{x}_i, \quad if\ F(\mathbf{x}_i) > F(\mathbf{x}_g^{best}) \tag{2.5}$$

which means $\mathbf{x}_i^{best}$ is the best position of particle i that had best fitness F in all iterations. Similarly, $\mathbf{x}_g^{best}$ is the best position obtained in all iterations among all particles in population.

- Update phase: At this step, based on the updated $\mathbf{x}_i^{best}$ and $\mathbf{x}_g^{best}$ at the previous step, particle velocity is updated by Eqn. 2.2. However, Eqn. 2.1 cannot be applied to update the particles' positions because of binary nature of the problem. Instead, the positions is calculated by the sigmoid function of the velocity as follows:

$$S(\mathbf{v}_i) = \frac{1}{1 + e^{-\mathbf{v}_i}} \tag{2.6}$$

The position of particle i can be updated by:

$$\mathbf{x}_i \longleftarrow \begin{cases} 0, & if\ Rand() > S(\mathbf{v}_i) \\ 1, & otherwise \end{cases} \tag{2.7}$$

where *Rand* is a random generator function that randomly selects a value from a uniform distribution in $[0, 1]$ range.

2.2.2 Feature Selection for GOs

Feature selection for the GO features sizing of over 10,000 is not practical with the Evolutionary Computation techniques. Thus a filter feature selection method, correlation-based measure is used to identify the highly correlated features. It means, features with greater predicting ability for a class are selected. We simultaneously eliminate the redundant features. Correlation-based feature selection method first calculates the correlation matrices of feature-class and feature-feature and searches the complete feature subset space. The main part of CFS is the heuristic evaluation of the selected feature subset described below (Hall, 2000)(Hall, 1999):

$$U_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2.8}$$

where U_S is a heuristic utility of a feature subset S which includes k feature, $\overline{r_{cf}}$ is the average feature-class correlation for all $f \in S$, and $\overline{r_{ff}}$ is the average inter-correlation between features in S. U is the utility metric that indicates the usefulness of every feature subset. The subset

with higher merit is used in order to reduce the dimensionality problem in the data. This reduced data set is then passed to machine learning scheme for training and testing purposes.

Exhaustive search for huge number of features (2^n for n feature) is not tractable as mentioned in Section 2.1.2. Therefore, to identify a feature subset with highest utility value, various search techniques have been used. Best first search was used as it is a highly effective search mechanism in CFS. Best first search method can start with either all features or with no features at all. Best first search with no features progresses in the search space by adding single feature at a time. When searching with all features, it moves backward through the search space by eliminating one feature at a time. The former was employed in this work. We initially used the set with no GO features. Then as the search moves forward, single GO feature is considered at a time. It can also backtrack and follow the backward path to a more promising previous subset in the feature space if the current search path is not found to be relevant. A stopping criterion is required to prevent the best first search from exploring the entire feature subset search space. The search converges if there is no improvement for the previous consecutive subsets compared to the current one.

There are different correlation metrics that can be used to compute correlations in Eqn. 2.8. Linear correlation techniques can be used to measure the correlation between two random variables (Yu and Liu, 2003). Even the non-linear methods using information theory can be used (Hall, 1999). Most popular correlation metrics used for CFS are Symmetric Uncertainty(SU), Minimum Description Length (MDL) and Relief (Hall, 1999). In this work, SU was implemented. SU is an entropy-based method. Entropy helps in capturing the purity of distribution of a random variable. For a random variable Y, entropy is calculated as:

$$H(Y) = -\sum_{y \in Y} p(y) log_2 p(y) \tag{2.9}$$

and after observing value of another variable X, the value of the entropy of Y given X is defined as:

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2 p(y|x) \tag{2.10}$$

where $p(x)$ is the prior probabilities for all values of X, and $p(y|x)$ is the likelihood probability. The amount of reduction in the entropy of X gives additional information about X is given by Y and is called Information Gain (IG) (Destrero et al., 2009), defined by:

$$IG(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(Y) + H(X) - H(X,Y) \tag{2.11}$$

IG should be normalized with their corresponding entropy values to favor the features with more information. Therefore, Symmetrical Uncertainty (SU) measure is used as follows:

$$SU(X,Y) = 2 \times \frac{IG(X,Y)}{H(X) + H(Y)} \tag{2.12}$$

To compute $\overline{r_{cf}}$ in Eqn. 2.8, X and Y in SU (Eqn. 2.12) are features ($f \in S$) and classes and similarly features and features to compute $\overline{r_{ff}}$.

Once the feature selection is done, to further reduce the size of features, we try to fit the data points provided by classifier into a polynomial function on the first important feature to all the features selected by CFS. Then, the top features are selected based on the knee point chosen by *Kneedle* method (Satopaa et al., 2011). The top number features can vary based on degree on fitted curve on the data points.

2.2.3 Classification

This work focuses on classifying if a chemical compound has a positive or negative impact on lifespan extension for a given values of features for the compound. This problem can be considered as a binary classification problem where 1 indicates an impact and 0 indicates the absence of an impact.

Each sample s_i in the data can be defined as $s_i = [\mathbf{f}_i, y_i]$ where $\mathbf{f} = [f_i^1, f_i^2, ..., f_i^d]$ is the d dimensional feature vector for sample s_i and y_i is the class label of s_i . The goal of classification here approximate a function $f : \mathbf{f} \rightarrow y$ to estimate the class label of a new sample s_j based on its feature vector $\mathbf{f}_j$. For comparison, we consider three classification methods, namely support vector machine, artificial neural network and random forest.

- **Support Vector Machine** (SVM) is a supervised machine learning algorithm for classification modeling. It frames a decision boundary for classifying all the training data while maximizing the margin between the classes. If there is no possibility for linear separation of data, then it uses a kernel function for realizing the non-linear mapping to new feature space. The hyper-plane in feature space found by SVM corresponds to non-linear boundary in the input space.

- **Artificial Neural Network** (ANN) is a parameterized supervised non-linear algorithm which consists of multi-layer perceptron with different number of perceptrons at each layer for the classification of input data. The numeric value of perceptrons, hidden layers in the model and the activation function used can decide the accuracy of the model.

- **Random Forest** (RF) is a well-known ensemble of decision trees and is widely used for classification and regression tasks. In RF, each tree contributes a vote for the assignment of most frequent class to input data (Breiman, 2001). When a tree grows, a random subset of features is selected for dividing a node with different bootstrap sample of data unlike the conventional decision tree methods. Thus, in a random forest, a node is split using the best among the subset of predictors which are randomly chosen at that node. Ultimately, the prediction is done by voting or averaging over all the trees.

2.3 Experimental Results

2.3.1 Data Set

From the DrugAge database (Barardo et al., 2017) found in Human Aging Genomic Resources website (Tacutu et al., 2012), the chemical compounds that improve the longevity on C. elegans were extracted. These compounds were considered as positive and were assigned a positive class label. Additionally, compounds that reduce or does not affect the lifespan were considered negative and were put under the negative class label. The data set consists of 229 positive samples and 1,166 negative samples.

Two types of features were used:

1. Molecular descriptors (MD): Chemical structure of a compound enables us to estimate a molecular descriptor. They help us in building the predictive models to study the relation between compounds chemical structure and its biological properties. The size of MD features is $N_{MD} = 268$ each of which takes up a numerical value. Among these 268 MDs, PSO feature selection algorithm selected a feature subset with 87 MDs. An example of molecular descriptor feature used in this work is a_nN. It defines the number of nitrogen atoms.

2. Gene Ontology (GO): the proteins that interact and get targeted by various compounds enables us in understanding the GO terms. They are classified into 3 categories: a) Biological Process b) Molecular Function and c) Cellular Components. The GO terms under biological process category are of utmost importance. Mitochondrial genome maintenance is an example of a biological process which defines about the maintenance of structure and integrity of mitochondrial genome including the replica and segregation of mitochondrial chromosome. Molecular Function GO terms reveal information about the molecular activities. A molecular function by name acyl binding describes the

activity of interacting selectively and non-covalently with an acyl group or any other group which is derived by removing the hydroxyl group from the acid function of carboxylic acid. The cellular component GO defines the location in the cell. Nuclear chromosome is a cellular component that is found in the nucleus of eukaryotic cell during the cell cycle phases when the nucleus is intact. It encodes the nuclear genome. An example of GO:0001941 is a biological process defined as a post synaptic membrane organization which is a process which results in the assembly, arrangement of constituent parts, or dis-assembly of a post synaptic membrane. In this work, a total of $13,338$ GO features were used. Thus, the size of GO features is $N_{GO} = 13,338$ Each feature takes a binary value. CFS selected a feature subset with 165 features. To further reduce the size of features, a polynomial function was used to fit the data points provided by the classifier accuracy by using one feature to 165 features. Then, top 16 features are selected based on the knee point chosen by Kneedle method (Satopaa et al., 2011). The number of features selected is inversely proportional to the degree of the polynomial function. Thus, number of features selected can be varied with the degree of fitting curve.

2.3.2 Performance Measurements

10-fold cross validation was implemented on DrugAge data. In this method, the data gets divided into 10 non-overlapping subsets and 9 among these 10 sets are used for training and the remaining subset is used for testing. Area Under the Curve (AUC) scores of the Receiver Operating Characteristic (ROC) are also reported in addition to the classification accuracy. As its the case of imbalanced data, AUC represents a better metric for performance estimation.

ROC represents a curve against True Positive Ratio (TPR) and False Positive Ratio (FPR) at different thresholds. The ways to estimate the accuracy, TPR and FPR are mentioned

below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.13}$$

$$TPR = \frac{TP}{TP + FN} \tag{2.14}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.15}$$

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively. The ratio of the positive data points that are correctly classified as positive with respect to all positive data points is the True Positive Ratio (TPR). False Positive Ratio (FPR) on the other hand denotes the ratio of negative data points that are wrongly considered as positive with respect to all negative points. The accuracy and AUC scores are obtained by averaging the results of 10-fold cross validation.

For PSO, we choose particle size as $P = 120$, maximum velocity $v_{max} = 0.9$ and maximum iteration $T = 150$. The inertia weight is $w = 0.9$, the local coefficient $c_1 = 2$ and global coefficient $c_2 = 2$. These values are taken from literature (Xue et al., 2013)(Moradi and Gholampour, 2016). In Eqn. 2.3, α is given a value 0.75 as classification performance is more important compared to number of features. But we chose $\alpha = 0.6$ as it gives a lower number of features with similar accuracy than with a value between (0.6, 1]. The initial position of particles is defined by (Vieira et al., 2013):

$$x_{ij} = \begin{cases} 1, & if\ Rand() > 0.5 \\ 0, & otherwise \end{cases}$$

Here *Rand* is a random generator function that selects values from $[0, 1]$ randomly. For CFS, a stopping criterion is imposed in best first search to avoid exploring the feature subset search space in its entirety. The search terminates if there is no considerable improvement observed on five consecutive, fully expanded subsets over the current best subset. Grid search and random search methods were used to find parameters for the classifiers to get a

better performance. Grid search checks for every combination of parameters and the one that maximizes the performance is taken into consideration. Randomized search on the other hand samples each setting from a distribution over possible parameter value. Radial Basis Function is used for SVM and for random forest. The maximum number of trees and features used in (Barardo et al., 2017) which are 900 and 210 are replaced with 300 and 152 as they gave better accuracy. ANN uses three hidden layers with 100, 300 and 100 neurons in each of the layers. Parameters are chosen based on the randomized search. Classification and wrapper feature selection methods were implemented on python 3. Random Forests and SVM were implemented using Scikit Learn and ANN was implemented using Tensorflow (Pedregosa et al., 2011). Filter Feature selection methods were implemented using Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009).

2.3.3 Results and Discussion

This section compares the performance of our feature selection method with the different available feature selection methods with a fixed classifier. PSO and Genetic Algorithm (GA) were combined to form a swarm-based wrapper for MD feature selection. CFS and Information Gain (IG) were combined to form a filter feature selection method for GO features. The classification results of mean of 10-fold cross validation in terms of accuracy and AUC are presented in Table 2.1. The results show an improvement in the accuracy and AUC score when used with the combination of PSO and CFS compared to other feature selection methods.

The performance of our feature selection method with different classifiers mentioned in section 2.2.3 was also analyzed and results are shown in Table 2.2. These results show that using the best features from our feature selection methods, the Random Forests perform better in terms of accuracy and AUC score compared to SVM and ANN.

The results were then compared with (Barardo et al., 2017) which uses Feature Importance (FI) for feature selection and classifies using a random forest classifier. The median of AUC

Table 2.1. Impact of feature selection on classification performance (RF classifier).

10-fold cross validation		
Methods	Accuracy	AUC-ROC
PSO+CFS	0.866	0.833
PSO+IG	0.827	0.805
GA+CFS	0.841	0.816
GA+IG	0.819	0.803

Table 2.2. Performance of classifiers for selected features.

10-fold cross validation		
Classifiers	Accuracy	AUC-ROC
ANN	0.792	0.713
SVM	0.801	0.702
RF	0.866	0.833

Table 2.3. Performance of classifier (RF) with different feature selections for three feature sets (MD, GO, MD+GO).

	FI+RF (previous)			PSO+CFS+RF (Ours)		
Performance	MD	GO	MD+GO	MD	GO	MD+GO
Accuracy	0.845	0.851	0.852	0.855	0.852	0.866
AUC-ROC	0.777	0.701	0.791	0.794	0.831	0.835

of 10-fold cross validation is shown as the performance measure in (Barardo et al., 2017). The mean of AUC and accuracy on 10-fold cross validation in (Barardo et al., 2017) and in our work were compared. For a proper comparison, we chose similar parameters for RF as in (Barardo et al., 2017). They included the number of trees and maximum number of features that can affect the performance of RF classification. To compute the performance of (Barardo et al., 2017), we picked the top 20 GO features and 73 MD features as the most important features and fed them to random forest classifier. Table 2.3 shows these results. It can be seen that MD features provide better AUC and accuracy over the GO features. Including both together slightly increases the performance. Table 2.3 shows the results which indicate an increase in the AUC and accuracy in our method. The AUC scores with standard deviation of 10 folds of cross validation along with the ROC curves of our approach, the

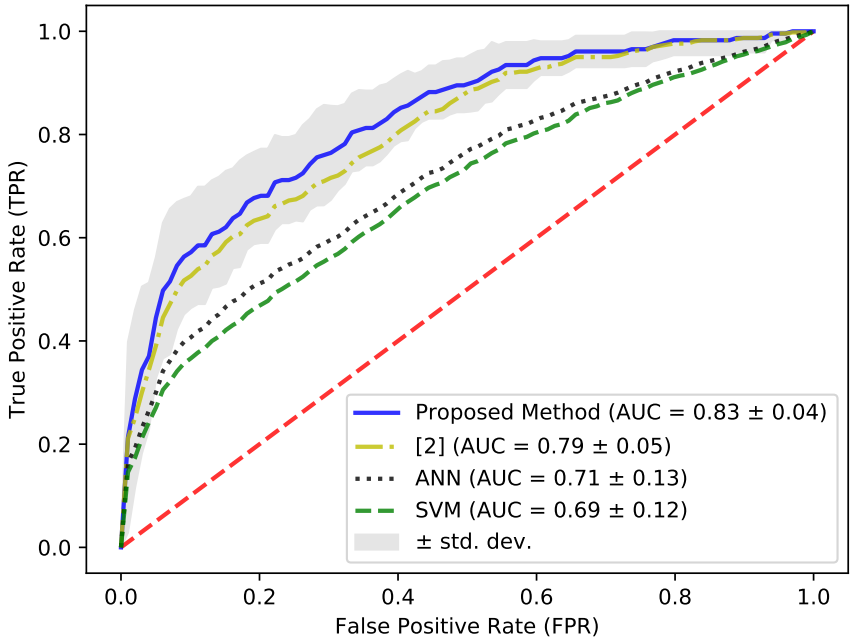

Figure 2.3. The ROC curves and AUC values for different methods.

approach in (Barardo et al., 2017), ANN and SVM classifiers are illustrated in Fig. 2.3. The grey area shows the standard deviation of our approach.

List of top 15 MD features (out of 268) provided by PSO and top 15 GO features (out of 165 features) selected by CFS, respectively, are provided in Table 2.4 and 2.5. In Table 2.5, the ontology (domain) of each GO term is represented by different colors. Red, green and blue colors represent, GOs belong to cellular component, molecular function and biological processes, respectively.

Table 2.4. Top 15 MD feature selected by Binary PSO.

Name	Description
a_nCl	Number of chlorine atoms
PEOE_VSA+4	Sum of vi where qi is in the range [0.20,0.25)
apol	Sum of the atomic polarizabilities
Q_RPC-	Relative negative partial charge
a_IC	Atom information content (total)
PEOE_VSA_FPPOS	Fractional positive polar van der Waals surface area
chi0v	Atomic valence connectivity index
SMR	Molecular refractivity (including implicit hydrogens)
a_nN	Number of nitrogen atoms
PEOE_PC-	Total negative partial charge
Q_VSA_HYD	Total hydrophobic van der Waals surface area
a_nBr	Number of bromine atoms
weinerPol	Wiener polarity number
ASA+	Water accessible surface area of all atoms with positive partial charge
a_nO	Number of oxygen atoms

Table 2.5. Top 15 GO feature selected by CFS.

Accession	Name
GO:0000164	protein phosphatase type 1 complex
GO:0000247	C-8 sterol isomerase activity
GO:0000506	GPI-GnT complex
GO:0001094	TFIID-class transcription factor binding
GO:0001510	RNA methylation
GO:0001546	preantral ovarian follicle growth
GO:0001869	negative regulation of complement activation
GO:0001941	postsynaptic membrane organization
GO:0002058	uracil binding
GO:0002059	thymine binding
GO:0002317	plasma cell differentiation
GO:0002762	negative regulation of myeloid leukocyte ...
GO:0003253	cardiac neural crest cell migration ...
GO:0003831	beta-N-acetylglucosaminylglycopeptide ...
GO:0003844	1,4-alpha-glucan branching enzyme activity

CHAPTER 3

WEISFEILER-LEHMAN NEURAL MACHINE FOR DRUG-TARGET LINK PREDICTION

Acknowledgement: The main part of this chapter will be utilized in the future work authored by Susmitha Sri Kadiyala, Eslami Manoochehri Hafez and Mehrdad Nourani.

3.1 Overview

Drug-Target Interactions:

Determining Drug-Target Interaction (DTI) is a major part in the area pharmaceutical sciences (Fakhraei et al., 2014). The process of drug discovery involves costs around $1.8 billion with a duration which may extend beyond 10 years (Ciociola et al., 2014). Thus, the drug-target interactions prediction helps in narrowing down the search area helping out the biologists. The main step in the process of drug discovery is to recognize the targets related to the drugs. These targets are mostly the proteins that can be drugged, and the ones related to diseases. The prediction of drug-target interactions aims at identifying the possible new targets for the existing drugs. It basically guides us through a proper experimentation process. There are many types of protein targets which include enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors. These classes can modulate their function by interacting with various ligands. Thus, analyzing the genomic space produced by these classes of proteins, helps us to accurately estimate the possibility of an interaction.

DTI can be used either for drug discovery or for re-positioning which is reusing available drugs for new targets. Approaches to predicting DTI can be classified into one of three categories: 1) Ligand-based, 2) Docking-based and 3) Chemogenomic approach. Prediction of DTI based on target proteins' ligand similarity is the Ligand-based approach (Keiser et al., 2007) (Keiser et al., 2009). 3D structure information of a target protein can help

in estimating the likelihood of its possible interaction with certain drug. This is based on their binding capacity and strength (Cheng et al., 2007) (Morris et al., 2009). This method is the Docking-based approach. Lastly, if the chemical information of the drugs, genomic information from proteins and already identified DTIs are used, then this is the chemogenomic approach (Mousavian and Masoudi-Nejad, 2014) (Ding et al., 2013).

A target with small number of binding ligands often leads to poor DTI predictions in a ligand-based method. This is a drawback in this method. Similarly, docking-based method is based on availability of 3D structures for target proteins and is also time consuming. These disadvantages made the chemogenomic approach more popular for identification of DTI in the recent times. This approach models the DTI problem as a machine learning problem and often builds a classifier which is trained by an available interaction data. This classifier is used to predict the unknown interactions (Ding et al., 2013).

Different techniques are employed in chemogenomic approach. Some of these include bipartite graph (Bleakley and Yamanishi, 2009)(Lu et al., 2017), recommendation systems (Alaimo et al., 2016) and supervised classification problem (Wen et al., 2017). But considering the data, we can see that there will be only a few positive interactions and the remaining possible interactions are unknown. For example, of the 35 million drug compound possible candidates, total number of positive drug interactions could just be 7000 (Bolton et al., 2008).

Computational chemogenomic approaches can be classified into feature-based and similarity-based methods. Feature-based methods have features as inputs for a set of instances defined by a particular class label. The instances are generally the drugs and the features are the targets. The class label is a binary value indicating the presence of a possible interaction. Some of the feature-based methods like decision trees, random forests (Breiman, 2001) and support vector machines (Cristianini and Shawe-Taylor, 2000) are used for classification purposes. Generally, Random Forest and Support Vector Machine are used for the classification of drug-target interactions (Yu et al., 2012).

Similarity-based methods take the similarity matrices of drugs and targets as inputs including the interaction matrix which indicates the drug-target pairs that most likely interact with each other. These approaches assume the common features among the drugs and targets can determine the presence of an interaction between a drug and target. The main focus is on the structural similarity between the drugs and targets leaving out the remaining unknown attributes. This gives an opportunity for easy implementation of the similarity indices on the networks without any prior information.

To implement the similarity-based algorithms, many similarity indices were used by (Lu et al., 2017). There are two ways to classify the similarity indices as local similarity indices and global similarity indices. The local similarity indices are node-dependent i.e. they require information related to neighbourhood of the network. On the other hand, global similarity indices are dependent on the path through the entire network topology (Lu et al., 2017). Use of the similarity indices outperform many random link predictors. A few examples of the similarity indices include Common Neighbours (CN), Jaccard Index, Preferential Attachment (PA) and Katz Index (Lu et al., 2017). These indices are described in the following:

- Common Neighbours: This defines the link by finding out if two nodes x and y share many common neighbours. If they do, then there is a high probability of having a link between them.

- Jaccard Index: This index measures the probability of having a common feature between two nodes x and y. The higher the probability, the higher is the possibility of a link between them. Its normalized version of common neighbours.

- Preferential Attachment: This is calculated according to the degree of nodes. The higher the degree, the more is the possibility for a link irrespective of the neighborhood of the nodes.

- Katz Index: This is a global similarity index and is therefore path dependent. It sums up the paths in the network and these paths are exponentially damped to result in shorter path. Longer the path, least is the Katz index value. It is basically a convergence method.

3.1.1 Prior Works

Drug-target link prediction is of high interest in the recent times. The application of machine learning in this challenging field of study makes it more promising and researchers have been trying to explore in depth in this field. (You et al., 2018) developed a lasso-based regularized linear classification method to predict the DTI overcoming the high-dimensional nature of the drug target data with huge number of features and small number of samples. A recent use of stacked auto encoder from the drug molecular structure and protein sequence to extract sufficient information was given by (Wang et al., 2018). DINES, a web server defined as drugtarget interaction network inference engine based on supervised analysis is used to predict unknown DTI for different biological data. It predicts with the help of machine learning methods integrating with the heterogeneous biological data in compatibility with the KEGG data (Yamanishi et al., 2014).

A framework was proposed by (Fakhraei et al., 2014) to work with a bipartite graph of drug-target interactions along with similarity measures. This used probabilistic soft logic to make predictions based on triad and tetrad structures. Enhancing the similarity measures to involve the non-structural information and to handle the possible missing interactions was done by (Shi et al., 2015). Matrix-factorization has been used in many research to identify new drug-target interactions (Gönen, 2012) (Eslami Manoochehri and Nourani, 2018) (Zheng et al., 2013). (Liu et al., 2016) used matrix-factorization method to focus on determining the probability of a drug-target interaction. In this method, the properties of drugs and targets were represented using their specific latent vectors. A modification to this method which

uses Bayesian optimization was developed by (Ban et al., 2017). It uses the neighborhood regularized logistic matrix factorization for DTI prediction. A multiple kernel link prediction on bipartite graphs was proposed by (Nascimento et al., 2016). In this method, relevant kernels are selected based on the weights which define the importance of a drug-target link.

Bipartite Local Model (BLM) was extended to DTI prediction with hubness-aware regression in (Buza and Peška, 2017). It builds a projection-based ensemble of BLMs using similarity space of drugs and targets. In (Lan et al., 2016), the unknown links are treated as unlabeled samples. A majority voting method is used to decide on the label of these unlabeled samples using the method of Positive-unlabeled learning for drugtarget prediction (PUDT). This unlabeled data is divided into reliable negative samples and likely negative samples with the help of their similarity information. Weighted SVM is used to then classify this data. We can also have positive unlabeled data similar to negative unlabeled data. To deal with this data, two group of approaches were proposed. One identifies the reliable negatives among the data and use the positives and these negatives to build the model. The other group directly predicts the positive unlabeled data (Liu et al., 2017).

Typically, there is a lot of unbalanced data in the drug-target interaction database. Weighted profile can be used to determine the drug profiles with defines weights as similarities between drug-drug (Yamanishi et al., 2008). The nearest profile is an approach that predicts the interaction profile of new drugs. (van Laarhoven and Marchiori, 2013) developed a simple weighted nearest neighbor procedure for predicting the interacting pairs with high accuracy. A network based inference helps in building a predictive model similar to a network where the nodes are represented by drugs and targets. Interacting drugs and targets are represented as edges (Cheng et al., 2012).

3.1.2 Key Contributions

Link prediction has found a noticeable place in many applications as social networking, e-commerce, etc. (Wang et al., 2015). Heuristic approaches have been mostly used for the

identification of any possible links. In the field of medicine, to predict a possible DTI, (van Laarhoven and Marchiori, 2013) has used common weighted nearest neighbors method. Some methods used shortest path analysis between the drugs and targets. These heuristic methods generally, do not reveal much information about the in-depth relations between drugs and targets.

In this work, a bipartite graph was first constructed with the available data. Then we worked on proper sampling of the huge number of negative samples in the data sets. We have used Weisfeiler-Lehman Neural Machine (WLNM) algorithm for creating the adjacency matrix which represents the interactions between potential drugs and targets. This adjacency matrix is fed to a machine learning model. We used 10-fold cross validation and AUC to estimate the efficiency. We compared the performance between Neural Network, Support Vector Machine and logistic regression models.

The chapter is organized as follows. We discuss our methodology in Section 3.2. It includes problem statement, pre-processing and the WLNM procedure. In Section 3.3 we present the experimental results of the work.

3.2 Methodology

3.2.1 Problem Statement

The interactions between drugs and targets can be represented using a bipartite graph. To attain higher stability with the efficiency in prediction of an interaction between drug and target, inclusion of drug-drug similarity and target-target similarity is beneficial. Moreover, for an imbalanced data with more negative samples, application of a sampling technique can help to deduct the unimportant data based on some scores. This helps a lot in improving the efficiency of the machine learning model.

We construct a bipartite graph for the prediction of Drug-Target link. The nodes in this graph represent the drugs and targets. The edges in this graph represent the links or

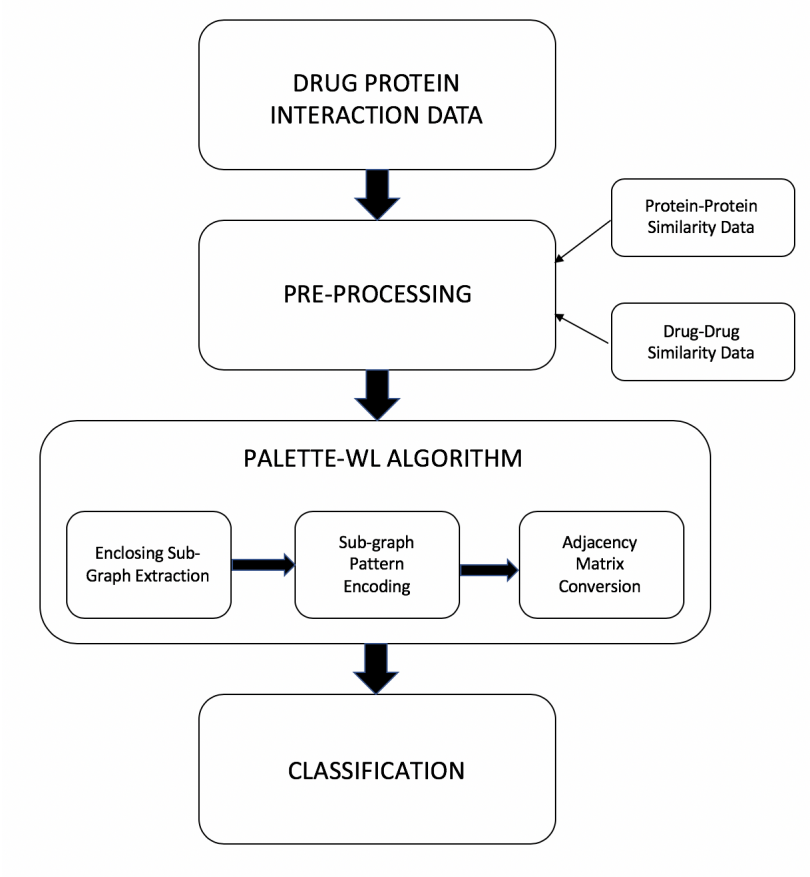

Figure 3.1. Work Flow of WLNM.

interactions between the drug and target. This interaction network can be represented using an adjacency matrix. Thus, for m drugs and n targets, $m \times n$ adjacency matrix X can represent the interactions as follows.

$$x_{ij} = \begin{cases} 1, & \text{if there is an interaction between } drug_i \text{ and } target_j \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

where x_{ij} denotes the $< i, j >$th element of matrix X $(1 \leq i \leq m, 1 \leq j \leq n)$. The aim of this work is to predict if x_{ij} is a possible interaction or not. The elements of the adjacency matrix with $x_{ij} = 0$ represent "unknown" interactions and $x_{ij} = 1$ represent the "positive" interactions i.e. drug d_i positively impacts target t_j.

3.2.2 Pre-processing

The main challenge of training a machine learning model for prediction of DTI is the data. Though the positive interactions (samples) between the drugs and targets are known, the negative samples representing the drugs and targets that surely do not interact with each other are not known. Most of the approaches in (Li et al., 2017) (Meng et al., 2017) (Wan et al., 2018) (Luo et al., 2017) and many more have selected the negative samples from the data set randomly. Random selection of negatives is not reliable. Study by Liu et. al. (Liu et al., 2015) proved that proper choice of negatives can greatly improve the performance of few major approaches like Bayesian Matrix Factorization (Gönen, 2012), bipartite local model (Bleakley and Yamanishi, 2009) and Gaussian kernel profile (van Laarhoven and Marchiori, 2013). In this work, we first identify reliable negatives using the following approach and then use the positives and the identified negatives for training the model.

Credible Negative Sampling: Using the converse negative proposition principle, we can say that a target which is dissimilar to set of predicted or known targets has higher probability of being dissimilar with a drug which is associated with an available interaction with this set of predicted or known targets. Similarly, a drug that is not similar to any of the predicted or known drugs that has interaction with a particular target, is less likely to interact with this target (Liu et al., 2015). These directives can be respectively called as *target dissimilarity* and *drug dissimilarity* rules. Using these two directives, we can find out the most reliable negatives among the possible DTIs. This method of negatives selection uses both the predicted and validated DTIs.

The process of identifying the negative samples among potential negatives is as follows:

- The similarity between a drug d_j and drug d_i is denoted by S_{ji}^D and the similarity between target k and target l is denoted by S_{kl}^T.
- The values of these similarities can be assumed as a combined score. They are added over the entire range for each drug and for each target (Liu et al., 2015). Let them be

denoted as S_j^D, S_k^T.

$$S_j^D = \sum_{i=1}^{m} S_{ji}^D \tag{3.2}$$

$$S_k^T = \sum_{l=1}^{n} S_{kl}^T \tag{3.3}$$

- The final score which is basically assumed to be the distance between drug and target is denoted as D_{jk} and is calculated using the formula below:

$$T = S_j^D + S_k^T \tag{3.4}$$

$$D_{jk} = e^{-T} \tag{3.5}$$

- Now rank the potential negatives according to the score D_{jk} in the decreasing order and those with highest values of the score are considered to be the potential negatives.

- The number of negatives is equal to the number of positives in our work.

- Once we have the positives and negatives, we can feed them to any classifier for classification purposes.

We used both random and credible methods for identification of negative samples in this work.

3.2.3 WLNM Process

The Weisfeiler-Lehman Neural machine (WLNM) is implemented for prediction of a Drug-Target Interaction. It follows three stages in accomplishing this task. These include extraction of the enclosing subgraph, pattern encoding of the extracted subgraph and neural network training. These steps are discussed below:

- **Extraction of Enclosing Subgraph:** For a deeper understanding of the, the WLNM algorithm extracts a subgraph for a drug-target link. The neighborhood's size is defined by the number of nodes in the enclosed subgraph. The number of nodes is defined by N given by the user. This subgraph defines the enclosing environment of the drug-target link. These topological details provided helps in determining the existence of a link. The process of extraction is as follows. For a given link $x - y$, the subgraph first starts adding the 1-hop neighbors into the node list V_N. Then, the hop is gradually increased to add the vertices into this list. Once the value of the number of vertices reaches N, the extraction of this subgraph stops (Zhang and Chen, 2017).

- **Pattern Encoding of the Extracted Subgraph:** The main step in the process of pattern encoding the subgraph is the graph-labeling part. Once the graph labeling is done, the subgraph chosen is made into an adjacency matrix. Then, this matrix is fed to a neural network in wlnm. Graph labeling should be properly and consistently done for getting a better accuracy in the algorithm. The vertex labeling using the graph labeling gives similar labels for the nodes with structural similarity. Using this along with maintaining the directionality in the topology for identifying the target link, a new graph labeling technique Palette WL algorithm was proposed by (Zhang and Chen, 2017).

  The normal WL graph labeling algorithm doesn't consider the directionality. These enclosed subgraphs developed by palette WL algorithm have the target link in the middle and the neighbors are added by iteration based on their distance to the link. The nodes close to the link get lower labels compared to the ones far away from the link. These vertices are sorted in ascending order and then nodes with redundant labels are removed with the help of a canonization tool called nauty (Zhang and Chen, 2017). Then, these subgraphs are represented as an upper triangular matrix and this adjacency matrix is fed to the classifier.

Table 3.1. Specifications of data sets

	IC	Enzyme	GPCR	NR
Drugs (D)	204	445	95	54
Targets (T)	210	664	223	26
Total DT Interactions	1476	2926	635	90

- **Training the Classifier:** As a final step, we trained a neural network with the positives and negatives represented by the enclosing adjacency matrix. The matrix is fed vertically to the feed forward neural network. We have also used SVM and Logistic Regression model along with the neural network.

3.3 Experimental Results

3.3.1 Data Sets

In this work, we used a public-domain data set (van Laarhoven and Marchiori, 2013) that corresponds to four different target protein types, namely nuclear receptors (NR), G protein-coupled receptors (GPCR), ION Channels (IC) and Enzymes (E). The number of drugs, targets and interaction among them is shown in Table 3.1. Each data set contains three matrices: $X \in \mathbb{R}_{m \times n}$ representing the DT interactions, $S^D \in \mathbb{R}_{m \times m}$ and $S^T \in \mathbb{R}_{n \times n}$ representing the similarities.

3.3.2 Performance Measurement

We have evaluated the given algorithm by starting with the pre-processing step. The data is inconsistent making it more challenging. Even though the positive samples are known, the negative samples representing the drugs and targets that do not interact with each other are not known. So, in this work, we used two methods for the selection of these negative samples. One is a random selection of negatives and the other is the credible sampling of the negatives.

Table 3.2. AUC for Random Sampling

Data Set	WL-NN	WL-LR	WL-SVM
IC	0.878	0.889	0.875
E	0.923	0.929	0.916
GPCR	0.805	0.858	0.823
NR	0.711	0.767	0.726

Table 3.3. AUC for Credible Sampling

Data Set	NN	LR	SVM
IC	0.972	0.929	0.935
E	0.966	0.943	0.935
GPCR	0.971	0.956	0.959
NR	0.869	0.951	0.932

Table 3.4. AUC for Similarity Based Methods

Data Set	CN	KI	JI	PA
IC	0.438	0.788	0.44	0.84
E	0.484	0.837	0.484	0.783
GPCR	0.47	0.75	0.473	0.759
NR	0.467	0.615	0.47	0.641

The AUC values obtained by using the Random Sampling method and Credible Sampling methods for the four data sets IC, GPCR, NR and E are shown in tables 3.2, 3.3, respectively.

Table 3.4 represents the AUC results for the standard similarity-based methods. We have evaluated this on four indices Jaccard Index (JI), Common Neighbours (CN), Preferential Attachment (PA) and Katz Index (KI) on the four data sets available.

3.3.3 Results and Discussion

The above experiments prove that using a proper sampling for gathering the reliable negatives highly improves the accuracy compared to the random selection of negative samples. Comparing the values of AUC for different models like Neural Network, SVM and Logistic Regression for the two negative sampling methods, we can infer that the AUC values have greatly improved by the credible sampling compared to the random selection method. Hence,

usage highly important unknowns for classification, improves the reliability of the classifier to a greater extent. Thus, we analyze the Weisfeiler-Lehman Neural Machine for learning the links in the network formed between the drugs and targets by constructing subgraphs.

Palette-WL algorithm was used to efficiently implement the graph labeling technique which considers the directionality of the nodes along with the structural similarity. Then, we built Neural Network, Logistic Regression and Support Vector Machine models for the training purpose for link prediction. Thus, the usage of proper negative sampling technique on the WLNM resulted in a better performance compared to the existing similarity-based methods.

CHAPTER 4

CONCLUSION AND FUTURE DIRECTIONS

The main focus of this research is on speeding up the target discovery stage of the drug discovery process. Application of machine learning in this stage eases up the process to a greater extent. We have proposed a model to predict compounds which are likely to improve the lifespan of C. elegans is created. This model is built using chemical molecular descriptors and gene ontology features. A feature subset was selected using a modified binary PSO algorithm for MD features and CFS method was used to select GO terms. Different classifiers were tested on the selected features and results show that the random forests achieve better results compared to SVM and ANN. This work can be extended by using some techniques which helps in dealing with the imbalanced data.

- As negative samples are high in number, they result in higher accuracy compared to positive samples. PSO has premature local minima. This performance can be improved by setting a proper value for inertia weight to balance its local search and global search. Classification performance for GO data can be enhanced using a different feature selection method.

- This method can be used to identify the unknown effect of drug compounds on lifespan on various other model organism like yeast, mouse and fly.

In the second work, we demonstrated that while predicting a possible link between a drug and target, consideration of highly important negative samples for training will result in construction of a highly efficient model. This DTI prediction helps the pharmaceutical industry in a very large scale. It helps in easing the process of drug discovery by a considerable factor. Usage of different subgraphing techniques in the algorithm proposed, will definitely improve the reliability of the training model. The following can be investigated as a future work:

- We can use the drug-drug and target-target similarities for the construction of subgraph and then implement a graph labeling technique for training a model.

- Moreover, this work is implemented on unbalanced data set as it has more negatives compared to the positives. In future we intend to apply few techniques to deal with this issue.

- Finally, this method can be applied to various data sets to predict the presence of any drug target interaction.

As a conclusion, identifying the aging effect of various drugs on the C. elegans can be extended to human beings due to genetic structural similarity between humans and C.elegans. Once the biochemists know about the drugs that improve the longevity, it will be helpful for the preparation of relevant medicines. On the other hand, efficient prediction of a drug-target link can definitely help in preparation of high-quality medicines. Thus, this research work adds many benefits in the field of drug discovery.

REFERENCES

Alaimo, S., R. Giugno, and A. Pulvirenti (2016). Recommendation techniques for drug–target interaction prediction and drug repositioning. *Data Mining Techniques for the Life Sciences*, 441–462.

Ban, T., M. Ohue, and Y. Akiyama (2017). Efficient hyperparameter optimization by using bayesian optimization for drug-target interaction prediction. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2017 IEEE 7th International Conference on*, pp. 1–6. IEEE.

Barardo, D., D. Thornton, H. Thoppil, M. Walsh, S. Sharifi, S. Ferreira, A. Anžič, M. Fernandes, P. Monteiro, T. Grum, et al. (2017). The drugage database of aging-related drugs. *Aging cell 16*(3), 594–597.

Barardo, D. G., D. Newby, D. Thornton, T. Ghafourian, J. P. de Magalhães, and A. A. Freitas (2017). Machine learning for predicting lifespan-extending chemical compounds. *Aging (Albany NY) 9*(7), 1721.

Batin, M., A. Turchin, S. Markov, A. Zhila, and D. Denkenberger (2017). Artificial intelligence in life extension: from deep learning to superintelligence. *Informatica 41*(4), 401–417.

Benchettara, N., R. Kanawati, and C. Rouveirol (2010). Supervised machine learning applied to link prediction in bipartite social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pp. 326–330. IEEE.

Bleakley, K. and Y. Yamanishi (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics 25*(18), 2397–2403.

Bolton, E. E., Y. Wang, P. A. Thiessen, and S. H. Bryant (2008). Pubchem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry 4*, 217–241.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Buffenstein, R., Y. H. Edrey, and P. L. Larsen (2008). Animal models in aging research. In *Sourcebook of Models for Biomedical Research*, pp. 499–506. Springer.

Burbidge, R., M. Trotter, B. Buxton, and S. Holden (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry 26*(1), 5–14.

Buza, K. and L. Peška (2017). Drug–target interaction prediction with bipartite local models and hubness-aware regression. *Neurocomputing 260*, 284–293.

Calvert, S., R. Tacutu, S. Sharifi, R. Teixeira, P. Ghosh, and J. P. Magalhães (2016). A network pharmacology approach reveals new candidate caloric restriction mimetics in c. elegans. *Aging Cell 15*(2), 256–266.

Carretero, M., R. L. Gomez-Amaro, and M. Petrascheck (2015). Pharmacological classes that extend lifespan of caenorhabditis elegans. *Frontiers in genetics 6*, 77.

Carretero, M., G. M. Solis, and M. Petrascheck (2017). C. elegans as model for drug discovery. *Curr Top Med Chem 17*, 1–10.

Cheng, A. C., R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, and E. S. Huang (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nature biotechnology 25*(1), 71–75.

Cheng, F., C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang (2012). Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology 8*(5), e1002503.

Ciociola, A. A., L. B. Cohen, P. Kulkarni, C. Kefalas, A. Buchman, C. Burke, T. Cain, J. Connor, E. D. Ehrenpreis, J. Fang, et al. (2014). How drugs are developed and approved by the fda: current process and future directions. *The American journal of gastroenterology 109*(5), 620–623.

Cristianini, N. and J. Shawe-Taylor (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dash, M. and H. Liu (1997). Feature selection for classification. *Intelligent data analysis 1*(3), 131–156.

Destrero, A., S. Mosci, C. De Mol, A. Verri, and F. Odone (2009). Feature selection for high-dimensional data. *Computational management science 6*(1), 25–40.

Ding, A.-J., S.-Q. Zheng, X.-B. Huang, T.-K. Xing, G.-S. Wu, H.-Y. Sun, S.-H. Qi, and H.-R. Luo (2017). Current perspective in the discovery of anti-aging agents from natural products. *Natural products and bioprospecting 7*(5), 335–404.

Ding, H., I. Takigawa, H. Mamitsuka, and S. Zhu (2013). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in bioinformatics 15*(5), 734–747.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM 55*(10), 78–87.

Eslami Manoochehri, H., S. S. Kadiyala, J. Birjandtalab, and M. Nourani (2018). Feature selection to predict compound's effect on aging. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 419–427. ACM.

Eslami Manoochehri, H. and M. Nourani (2018). Predicting drug-target interaction using deep matrix factorization. In *Biomedical Circuits and Systems Conference (BioCAS), 2018 IEEE*, pp. 1–4. IEEE.

Fabris, F., J. P. De Magalhães, and A. A. Freitas (2017). A review of supervised machine learning applied to ageing research. *Biogerontology 18*(2), 171–188.

Fabris, F., A. Doherty, D. Palmer, J. P. de Magalhães, A. A. Freitas, and J. Wren (2018). A new approach for interpreting random forest models and its application to the biology of ageing. *Bioinformatics 1*, 8.

Fakhraei, S., B. Huang, L. Raschid, and L. Getoor (2014). Network-based drug-target interaction prediction with probabilistic soft logic. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 11*(5), 775–787.

Fontana, L., L. Partridge, and V. D. Longo (2010). Extending healthy life spanfrom yeast to humans. *science 328*(5976), 321–326.

Gasca, E., J. S. Sánchez, and R. Alonso (2006). Eliminating redundancy and irrelevance using a new mlp-based feature selection method. *Pattern Recognition 39*(2), 313–315.

Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics 28*(18), 2304–2310.

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio (2016). *Deep learning*, Volume 1. MIT press Cambridge.

Guzella, T. S. and W. M. Caminhas (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications 36*(7), 10206–10222.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Hall, M. A. (1999). Correlation-based feature selection for machine learning.

Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

Kaletta, T. and M. O. Hengartner (2006). Finding function in novel targets: C. elegans as a model organism. *Nature reviews Drug discovery 5*(5), 387.

Keiser, M. J., B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet (2007). Relating protein pharmacology by ligand chemistry. *Nature biotechnology 25*(2), 197–206.

Keiser, M. J., V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, et al. (2009). Predicting new molecular targets for known drugs. *Nature 462*(7270), 175–181.

Kennedy, J. (2011). Particle swarm optimization. In *Encyclopedia of machine learning*, pp. 760–766. Springer.

Lan, W., J. Wang, M. Li, J. Liu, Y. Li, F.-X. Wu, and Y. Pan (2016). Predicting drug–target interaction using positive-unlabeled learning. *Neurocomputing 206*, 50–57.

Learning, M. (2012). Tom mitchell. *Machine Learning 10*, 601.

Li, Z., P. Han, Z.-H. You, X. Li, Y. Zhang, H. Yu, R. Nie, and X. Chen (2017). In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Scientific Reports 7*(1), 11174.

Liu, H., J. Sun, J. Guan, J. Zheng, and S. Zhou (2015). Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics 31*(12), i221–i229.

Liu, Y., S. Qiu, P. Zhang, P. Gong, F. Wang, G. Xue, and J. Ye (2017). Computational drug discovery with dyadic positive-unlabeled learning. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 45–53. SIAM.

Liu, Y., M. Wu, C. Miao, P. Zhao, and X.-L. Li (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS computational biology 12*(2), e1004760.

Lu, Y., Y. Guo, and A. Korhonen (2017). Link prediction in drug-target interactions network using similarity indices. *BMC bioinformatics 18*(1), 39.

Lucanic, M., G. J. Lithgow, and S. Alavez (2013). Pharmacological lifespan extension of invertebrates. *Ageing research reviews 12*(1), 445–458.

Luo, Y., X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications 8*(1), 573.

Mao, Q. and I. W.-H. Tsang (2013). A feature selection method for multivariate performance measures. *IEEE transactions on pattern analysis and machine intelligence 35*(9), 2051–2063.

Meng, F.-R., Z.-H. You, X. Chen, Y. Zhou, and J.-Y. An (2017). Prediction of drug–target interaction networks from the integration of protein sequences and drug chemical structures. *Molecules 22*(7), 1119.

Moradi, P. and M. Gholampour (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing 43*, 117–130.

Moradi, P. and M. Rostami (2015). A graph theoretic approach for unsupervised feature selection. *Engineering Applications of Artificial Intelligence 44*, 33–45.

Morris, G. M., R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson (2009). Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry 30*(16), 2785–2791.

Mousavian, Z. and A. Masoudi-Nejad (2014). Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert opinion on drug metabolism & toxicology 10*(9), 1273–1287.

Nascimento, A. C., R. B. Prudêncio, and I. G. Costa (2016). A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics 17*(1), 46.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Putin, E., P. Mamoshina, A. Aliper, M. Korzinkin, A. Moskalev, A. Kolosov, A. Ostrovskiy, C. Cantor, J. Vijg, and A. Zhavoronkov (2016). Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging (Albany NY) 8*(5), 1021.

Satopaa, V., J. Albrecht, D. Irwin, and B. Raghavan (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, pp. 166–171. IEEE.

Shi, J.-Y., S.-M. Yiu, Y. Li, H. C. Leung, and F. Y. Chin (2015). Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods 83*, 98–104.

Shi, Y. and R. Eberhart (1998). A modified particle swarm optimizer. In *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, pp. 69–73. IEEE.

Siedlecki, W. and J. Sklansky (1993). A note on genetic algorithms for large-scale feature selection. In *Handbook Of Pattern Recognition And Computer Vision*, pp. 88–107. World Scientific.

Tabakhi, S., P. Moradi, and F. Akhlaghian (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence 32*, 112–123.

Tacutu, R., T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. Costa, V. E. Fraifeld, and J. P. De Magalhaes (2012). Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic acids research 41*(D1), D1027–D1033.

van Laarhoven, T. and E. Marchiori (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS one 8*(6), e66952.

Vieira, S. M., L. F. Mendonça, G. J. Farinha, and J. M. Sousa (2013). Modified binary pso for feature selection using svm applied to mortality prediction of septic patients. *Applied Soft Computing 13*(8), 3494–3504.

Wan, F., L. Hong, A. Xiao, T. Jiang, and J. Zeng (2018). Neodti: Neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *bioRxiv*, 261396.

Wang, L., Z.-H. You, X. Chen, S.-X. Xia, F. Liu, X. Yan, Y. Zhou, and K.-J. Song (2018). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *Journal of Computational Biology 25*(3), 361–373.

Wang, L., N. Zhou, and F. Chu (2008). A general wrapper approach to selection of class-dependent features. *IEEE Transactions on Neural Networks 19*(7), 1267–1278.

Wang, P., B. Xu, Y. Wu, and X. Zhou (2015). Link prediction in social networks: the state-of-the-art. *Science China Information Sciences 58*(1), 1–38.

Wen, M., Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, and H. Lu (2017). Deep-learning-based drug–target interaction prediction. *Journal of Proteome Research 16*(4), 1401–1409.

Xue, B., M. Zhang, and W. N. Browne (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE transactions on cybernetics 43*(6), 1656–1671.

Xue, B., M. Zhang, W. N. Browne, and X. Yao (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation 20*(4), 606–626.

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

Yamanishi, Y., M. Kotera, Y. Moriya, R. Sawada, M. Kanehisa, and S. Goto (2014). Dinies: drug–target interaction network inference engine based on supervised analysis. *Nucleic acids research 42*(W1), W39–W45.

Ye, X., J. M. Linton, N. J. Schork, L. B. Buck, and M. Petrascheck (2014). A pharmacological network for lifespan extension in caenorhabditis elegans. *Aging cell 13*(2), 206–215.

You, J., M. M. Islam, L. Grenier, Q. Kuang, R. D. McLeod, and P. Hu (2018). Drug-target interaction network predictions for drug repurposing using lasso-based regularized linear classification model. In *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*, pp. 272–278. Springer.

Yu, H., J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang, and Y. Wang (2012). A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PloS one 7*(5), e37608.

Yu, L. and H. Liu (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863.

Zernov, V. V., K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev (2003). Drug discovery using support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of chemical information and computer sciences 43*(6), 2048–2056.

Zhang, M. and Y. Chen (2017). Weisfeiler-lehman neural machine for link prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 575–583. ACM.

Zhang, M.-L. and Z.-H. Zhou (2014). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering 26*(8), 1819–1837.

Zheng, X., H. Ding, H. Mamitsuka, and S. Zhu (2013). Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1025–1033. ACM.

Ziehm, M., S. Kaur, D. K. Ivanov, P. J. Ballester, D. Marcus, L. Partridge, and J. M. Thornton (2017). Drug repurposing for aging research using model organisms. *Aging cell 16*(5), 1006–1015.

BIOGRAPHICAL SKETCH

Susmitha Sri Kadiyala was born in Andhra Pradesh, India. She is the eldest daughter of Kadiyala Venkata Ramanjaneyulu and Kadiyala Nagamani Kumari. She completed her Bachelors in Technology in Electronics and Communication Engineering from G. Narayanamma Institute of Technology and Science, Hyderabad in 2016. With a good academic standing, Susmitha was awarded with the Young Engineer Award during her Bachelors. She regularly coordinated and organized many activities under the NGO Street Cause back in India. She was admitted to the Master's program in Computer Engineering at The University of Texas at Dallas in August 2016 . She was awarded the Jonsson School Graduate Scholarship for year 2016-2017 and Pathways to Research Scholarship for the academic year 2017-2018.

CURRICULUM VITAE

Susmitha Sri Kadiyala

November 19, 2018

Contact Information:

Department of Electrical and Computer Engineering
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080-3021, U.S.A.
Voice: (682) 256-2868
Email: susmithasri.kadiyala@utdallas.edu

Educational History:

B.Tech., Electronics and Communication Engineering, G.Narayanamma Institute of Technology and Science, 2016
M.S., Computer Engineering, University of Texas at Dallas, 2018

Application of Machine Learning in Drug Discovery
M.S. Thesis
Electrical and Computer Engineering Department, University of Texas at Dallas
Advisors: Dr. Mehrdad Nourani

Employment History:

Software Developer Intern, IBM, Austin, August 2017 – present

Professional Recognitions and Honors:

Erik Jonsson School Graduate Study Scholarship for the academic year 2016-2017
Pathways to Research scholarship for the academic year 2017-2018
Young Engineer Award for the academic year 2015-2016

Relevant Coursework:

Computer Architecture
Discrete Structures
Machine learning
Database Design
Design Analysis of Algorithms
Advanced Operating Systems

Research Interests:

Bio-Mechanics
Bio-Informatics
Machine Learning