Detecting Adverse Drug Reaction (ADR) with Data Mining and predicting its intensity with Machine Learning

SUBMISSION DATE: 22.07.18

Submitted by:

Nadib Hussain (15101080)
Tanvir Islam (15101113)
Rafik Un Nabi Apu (13101054)
Department of Computer Science and Engineering

Supervisor:

Amitabha Chakrabarty, Ph.D

Assistant Professor

Department of Computer Science and Engineering

Declaration

We, hereby declare that this thesis is based on results of our own findings. Materials of work from researchers conducted by others are mentioned in the references. This thesis, neither in parts nor as a whole, have been submitted previously by anyone of any institute or university for the award of any degree.

Signature of Supervisor

Amitabha Chakrabarty, Ph.D

Assistant Professor

Department of Computer Science and Engineering

BRAC University

Signature of Authors

Nadib Hussain (15101080)

Tanvir Islam (15101113)

Rafik Un Nabi Apu (13101054)

ABSTRACT

Adverse Drug Reaction (ADR) is one of the many uncertainties which are considered as a fatal threat in the field of pharmacy and medical diagnosis. Utmost care is taken to test a new drug thoroughly before it is introduced and made available to the public; but these pre-clinical trials are not enough on their own to ensure safety. Many ADRs are discovered in the later stages of consumption which could not be found out during the pre-clinical trials. The increasing concern to the ADRs has motivated the development of statistical, data mining and machine leaning methods to detect the Adverse Drug Reactions. With the availability of electronic health Records (EHRs) it has become possible to detect ADRs with the mentioned technologies. In this work, we have proposed a hybrid model of data mining and machine learning to identify different Adverse Reactions and predict the intensity of the final outcome. We have used the Proportionality Reporting Ratio (PRR) along with the precision point estimator test called the Chi-Square test to mine the different associations between drug and symptoms called the drug-ADR association. This output from the data mining technique is used as an input to the machine learning algorithms such as Random Forest and Support Vector Machine (SVM) to predict the intensity of the final outcome of ADR, depending on a patient's demographic data such as gender, weight, age, etc. We have performed the analysis on a total count of 88000 data taken from the publicly available dataset of FDA and achieved an accuracy of 91% to predict 'death' as the final outcome from an ADR.

Keywords: Adverse Drug Events, Healthcare, Medical Diagnosis, Data mining, Machine Learning, Random Forest, Support Vector Machine. Drug- Symptom association.

Acknowledgement

As this is the first time we work on data science, we were unfamiliar with most of the techniques and methods. However, we got huge support finding the right data and selecting the right topic from our supervisor Amitabha Chakrabarty. Moreover, as we knew very little about machine learning and data mining, our respected faculties like Dr. Mahbub Alam Majumdar and Samiul Islam guided us with a lot of patience and helpful knowledge. More over in our survey we took help from Dr. Tuhin Sultana and Dr. A. Alam Chowdhury, lecturers of Bangabandhu Sheikh Mujib Medical University.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

In this report we made an attempt to discuss various fields of data science and machine learning. We have used open-source dataset on Drug reaction of FDA (2014-2017) [1]. We have designed a hybrid model of Machine learning and data mining through which we can predict different kinds of adverse drug reactions and its intensity on the basis of the user's demographic characteristics. In our research we have used various algorithms like Support Vector Machine (SVM), Random forest & Neural Network.

1.1 Motivation

Nowadays almost 73% of the people around the world take different medications. Among these, almost 29% [19] of these medicines have different kinds of adverse drug reactions. FDA statistics shows that almost 7000 [12] of these ADRs have caused death in recent years. So, through this research we tried to predict the ADR and its intensity so that necessary precautions can be taken before prescribing any kinds of medications.

1.2 Objective

The main purpose of our research is to use various tools of data science and machine learning to reduce casualties from ADR. Moreover, by using our hybrid model we are trying to predict ADR so that doctors can get assistance while prescribing medicines to their patients. In our research we used various algorithms and demonstrated how each these algorithms perform on the FDA dataset.

Through this research we can also help other future works by showing how these algorithms react on different features of the dataset. We also tried to show how much each of the features are relevant with each other. So, by getting values from our research, future researchers can easily decide which algorithm would be best to use for these kinds dataset.

1.3 Methodology

In our hybrid model we have two separate parts, first part is the data mining model which uses statistical methods like Proportionality Reporting Ratio (PRR) and Chi-Square. In this part we use certain threshold to estimate possible ADR that may affect the user. The input of the data mining model would be patient's symptoms and the medicines that the doctor in prescribing. The output of this model would be the possible ADRs which has a higher score of PRR or Chi-Square then our threshold.

The output of our first model will be included as input of our second model with some additional demographic information of the patient taken from the dataset. In the second part of our system we will use machine learning to predict the intensity of the ADR. In this part we will take various demographic information of the patients like age, gender, weight etc. and use them as training features in our machine learning model to predict the intensity of the ADR.

1.4 Problem Statement

In this research the biggest problem we faced was the dataset itself. As mentioned before we used the open source FDA dataset. Though the size of the data was quite huge, however it can't be considered as an ideal dataset for any kind of machine learning algorithms. It had a lot of null value. Moreover, all the features available in the data set have very poor relevance to the outcome.

One of the main problems of the dataset is that most of the features available is non-numerical categorical data. So, we had to use various methods like Vectorization, One-hot encoding, label encoding etc. to handle these kinds of data. These methods will be discussed later in the report.

Detecting ADRs is very hard even for the professional doctors. Different ADR can occur for same reason to different persons. On the other hand, same ADR can happen from two totally different reasons. So, predicting ADR can be very tough even for the doctors.

1.5 Solutions

We used various method to handle the FDA dataset. We used many libraries from python like scikit-learn etc. to handle null values, redundant data, categorical data etc. so that it can be fed into our model to train it.

Moreover, we consulted two professors of Bangabandhu Sheikh Mujib Medical University for selecting the features and understanding how much these features are relevant with the outcome.

1.6 Thesis Orientation

Chapter 1 is INTRODUCTION. The Motivation and Objectives of the thesis are described here.

Chapter 2 is LITERATURE REVIEW. This chapter consists of “Background” which defines the problem space of our thesis. Again, “Literature Review “indicates our information collection repository. This chapter also consists of “Supervised Learning” which is a brief description of how we initiated our system.

Chapter 3 is METHODOLOGY, where a description of our dataset can be found. Moreover, this chapter also contains “Implementation” where three algorithms of Machine Learning are described.

Chapter 4 is RESULT ANALYSIS, where we have shown how different algorithms performed with respect to our dataset.

Chapter 5 is CONCLUSION & FUTURE WORK consisting “Conclusion” and “Future Work".

CHAPTER 2

Literature Review

This chapter will contain the background research and related existing work summaries.

2.1 Background Study

Adverse drug Reactions (ADR) are injuries or harmful effects caused by a drug or the use of a drug [1]. Occurrences of ADR has increased largely over the past few decades. Worldwide, around 4.9% of hospital admissions are the result of ADEs and this number is as high as 41.3% in some areas [2]. In Sweden, ADRs are the seventh most common cause of death [3]. Even though drugs are thoroughly tested clinically before they are released on the market, many unknown side effects are discovered after they have been used over time by various patients. But point to be noted: a wrongful overdose of drugs is not considered as an ADR case.

2.1.1 Related Works

The aforementioned challenges motivated a series of works that apply data mining and machine learning approaches to come up with various solutions using different datasets to suit according to researcher's needs. Google recently worked on a Twitter data set to detect Adverse Drug Effects (ADEs) from posts on Twitter using a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [4]. They used a binary classifier to represent the outcome. Another researcher proposed a Predictive Pharmacosafety Networks (PPNs) to detect likely unknown ADEs [5]. He used the existing drug safety information from a well-known data set of drug

safety of 2005 to train a logistic regression model to detect unknown ADEs. In 2012 a research used the THIN database to create a model using feature matrix and feature selection to identify ADRs for a specific drug called “Pravastatin” [6]. A research in 2017 proposed a model using Electronic health records and train them in Random Forest algorithm to predict future unknown adverse drug reactions [7].

The random forest algorithm uses a decision tree approach to create a forest and select the outcome with the maximum count as their output. Here[8]a research paper showed a data mining technique called “Casual Association Rule” to find a cause-effect relationship between the drug-symptom pair which can be used to detect ADRs. In 2013another researcher named Duan proposed two novel models - a likelihood ratio model and a Bayesian network model—to create a pattern discovery method to identify adverse drug reactions [9]. A new research [10]used data mining techniques to find a relationship between a combination of drugs and their possible ADRs using the Chi-Square and Proportionality Reporting Ratio (PRR) as their basis to find the relationship. Our research depends heavily on this idea as we have used similar data mining techniques for our work.

2.1.2 Supervised Learning

Maximum practical machine learning algorithms uses supervised learning. In supervised learning we have input variable(X), output variable (y) and then we use an algorithm to learn the mapping function from input and output.

$$Y = f(X) \quad (1)$$

The goal is to approximately map the function so well that when there is new input data (x), the algorithm can predict the output variables (y) for that data. It is called supervised learning because the learning process of an algorithm from the training dataset can be thought of as a teacher supervising the learning process, where the teacher knows the correct answers; the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance [11]

CHAPTER 3

Methodology

As mentioned before we have proposed a hybrid model of two different technologies. Our whole system has two separate parts.

1. Data mining
2. Machine learning

Before we started our work, we had to manage the dataset, filter it and format it so that it can be used in our system.

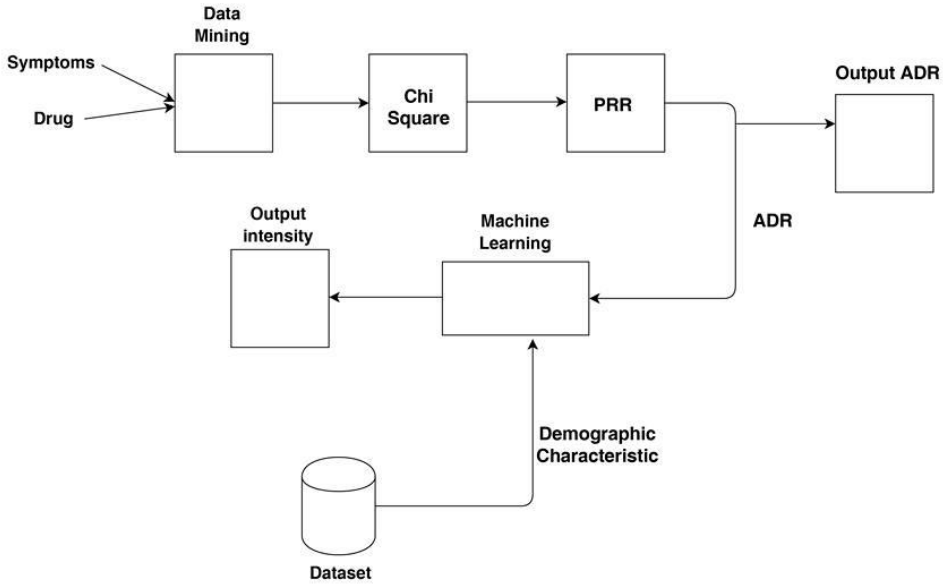

Figure 3.1 Hybrid System

3.1 Dataset

We have used an open-source data set that was published by the Food and Drug Administration(FDA). FDA publishes a dataset every 4 month of the year. FDA publishes data both in CSV and ASCII format.

3.1.1 FDA

Food and Drug Administration (FDA) is a federal agency of the United States Department of Health and Human Services established at June 30, 1906. The Food and Drug Administration ensure the safety, efficacy and security of human veterinary drugs, biological products, and medical devices to protect the public health. And FDA is responsible for advancing the public health by helping to speed innovations that make medical products more effective, safer, and more affordable and by helping the public get the accurate, science-based information they need to use medical products and foods to maintain and improve their health. FDA also plays significant role in maintaining food product's quality and to reduce tobacco uses by minors.

FDA has database for approved drugs. Information about FDA approved brand name and generic prescription and over the counter human drugs and biological products. FDA includes most of the drug products approved since 1939. The majority of the patient's information, labels, approval letters, reviews, and other information are also available for drug products approved since 1998 [14].

3.1.2 Data

We used the ASCII format of the data. In the dataset there were 7 separate tables. Though each of the tables have their own primary key, mainly all of the tables are

connected with the demographic table. All the table have 1 to many relations with the demographic table.

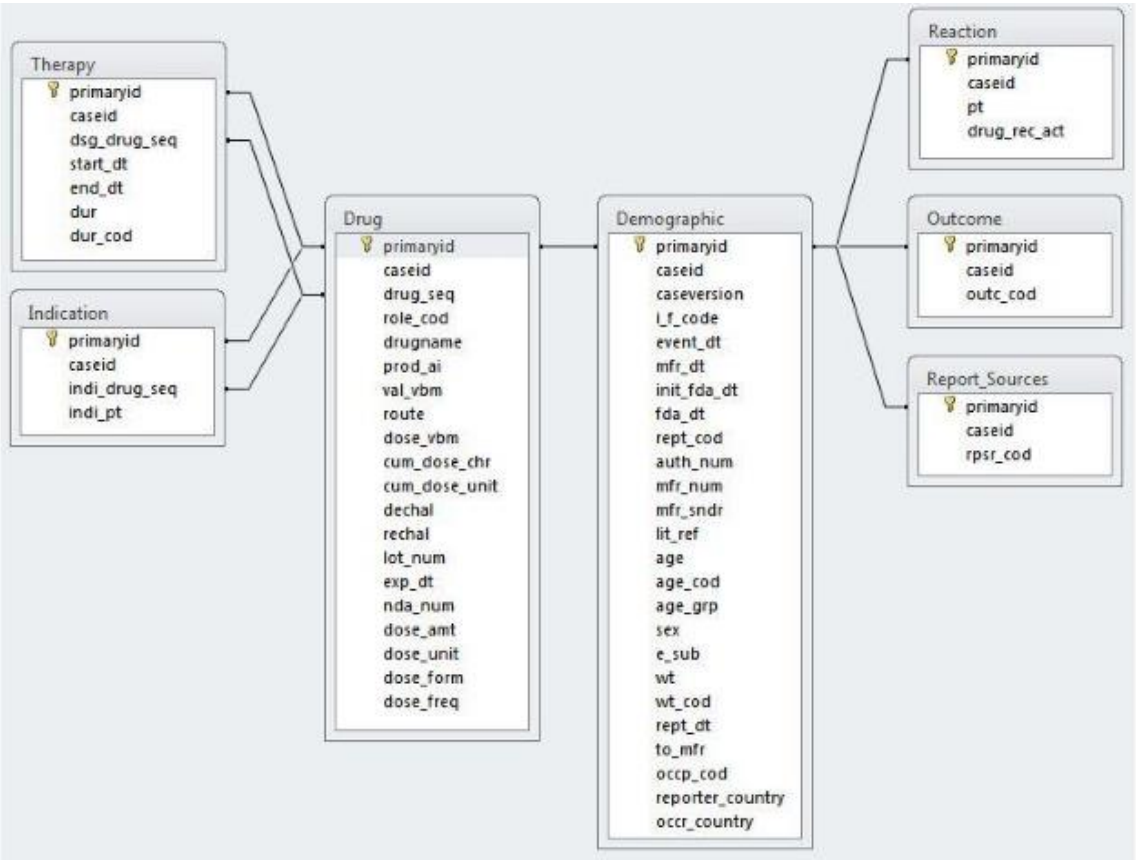

Figure 3.2 Dataset

We used Microsoft Access to filter out and convert the data to CSV format. For initial filtering we used some basic built-in filters from Microsoft Access to format out the initial data. Though the amount of data was huge (17 million Electronic Records). Most of them had noises. Many of the data were corrupted and most of the features of the data were null. We had to filter the whole data very carefully because these kinds of null and noise could badly impact our whole model and create large outliers in our results.

3.2 Data-Mining Model

Our System starts with the data mining part of the model. In [18] [1] we have seen that some statistical models like Chi-Square and PRR can be used to predict ADR. In this model user will input the **Symptoms** and the **Drug Names.**

3.2.1 Chi Square & PRR

Table 3.1 Variable of model

	Event	**All Other Events**	**Total**
Medical Product	**A**	**B**	**A+B**
All Other Product	**C**	**D**	**C+D**
Total	**A+C**	**B+D**	**N=A+B+C+D**

$$\mathbf{PRR}=\frac{\boldsymbol{A}/(\boldsymbol{A}+\boldsymbol{B})}{\boldsymbol{C}/(\boldsymbol{C}+\boldsymbol{D})} \quad (2)$$

The general criteria to run the PRR are as follows:

- Value A indicates the number of individual cases with the suspect medicinal product P involving an adverse event R.
- Value B indicates the number of individual cases related to the suspect medicinal product P, involving any other adverse events but R.
- Value C indicates the number of individual cases involving event R in relation to any other medicinal products but P.
- Value D indicates the number of individual cases involving any other adverse events but R and any other medicinal products but P

$$\mathbf{ChiSquare}, \mathbf{x^2} = (\mathbf{AD} - \mathbf{BC})^{\mathbf{2}} + (\mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D})/[(\mathbf{A} + \mathbf{B})(\mathbf{C} + \mathbf{D})(\mathbf{A} + \mathbf{C})(\mathbf{B} + \mathbf{D}) \quad (3)$$

3.2.2 ADR Detection

When the PRR is displayed with chi square and ADR is detected:

- PRR > 2
- Chi Square >4
- the number of individual cases greater or equal to 3

We calculated all the values in python's panda libraries. We inserted 2 more columns into the data frame [20]. After calculating all the variables needed for the equation we inserted

the corresponding row to see if that ADR falls under any of the conditions. Then these ADRs are matched with already given symptoms by the user. If there is any cross match, then that is considered as a potential ADR.

drugname	pt	chiSquare	PRR
SALINE	Musculoskeletal stiffness	1.275823	0.000000
DECADRON	Musculoskeletal stiffness	1.275823	0.000000
GASTER	Musculoskeletal stiffness	1.275823	0.000000
TAXOL	Musculoskeletal stiffness	1.275823	0.000000
DIPHENHYDRAMINE HCL	Musculoskeletal stiffness	1.275823	0.000000
ABILIFY	Agitation	0.918084	0.000000
CLOZAPINE.	Agitation	0.918084	0.000000
VALPROIC ACID.	Agitation	0.918084	0.000000
ATENOLOL.	Agitation	0.918084	0.000000
ABILIFY	Anxiety	20.521441	3.579423

Figure 3.3 Data Mining results

The output of this model is then inserted in to our second model which is the machine learning model.

3.3 Machine Learning Models

In this part of our system we are trying to predict the intensity of the ADR that was the output of the data mining model. Before we can use any kind of machine learning model we had to filter our data again and make some important changes so that we could feed our data to the model and make the predictions.

3.3.1 Feature Selectionwith Pearson Corelation Coeffitient

There were almost 30 features availabe in the dataset.Selecting proper feature is one of the main part of machine learning.Most of these features were irrelevant with our model and problems.We had to get all the tables together by using their primary Id.After getting all the features together [15], we have used Pearson's co-relation method for getting the best features.Though the corelation coefficients were very low,we took the top features which had the highest values of the corelation coefficient.

3.3.2 Feature Selection with Doctors opinions

As mentioned above, the FDA dataset has 7 tables with a lot of features. To find out which features are most important to the outcome, we went to a few doctors who are quite adept in the field of ADRs, gave them a list of the features from all the tables and ask them to give weights to the features on the basis of their importance to the outcome within a scale of 1 to 10. We consulted with 2 Professors of Bangabandhu Sheikh Mujib Medical University.

3.3.3 Data Management

We created our dataset on the combined basis of the Pearson's correlation and the survey that we did. After we have our very own data set with 7 filtered out features, we had to shape dataset according to our own needs. Here are all the steps that we have used to shape the dataset:

- Remove any null values from any of the features.
- Convert the categorical feature columns into numerical value for machine learning algorithms to understand.

- For categories with few unique values, we used pandas “get-dummies” method to One-Hot encode them into numerical values. For example the gender feature is converted into numerical value the same way [16].
- We have a categorical column called “pt” which describes the adverse reaction that occurred. It has more than 5000 unique values which makes it very difficult to meaningfully convert all the values into numeric.
- We generalized the “pt” column on the basis of specific body parts, for example: ‘skin’, ‘heart’, ‘Abdominal Pain’, etc. and then selected 32 most occurring adverse reactions to keep in our dataset.
- We then used “get-dummies” method to convert the categorical values into numeric values.[17]
- Another feature is called “route” which describes the route of the drug intake. It had 64 unique values, so again we reduced it to 14 highest occurring values and then converted it to numerical values using “get-dummies”.
- The label of our dataset that is the outcome that we want to predict is also a categorical value. The outcome column has 7 different classes of values. We generalized the classes into 3 different categories, which are – “Death”, “Serious Injury”, and “Minor Injury”.
- Using these three categories we used “Label encoding” to encode them to numerical values.
- We used Python dictionaries to generalize data.

abc= {“out_cod”: {“HO:1, “DE”:2, “LT”:1, “OT”:0, “DS”:0, “CA”:0, “RI”:0,}

df.replace (abc, inplace=True)}

Finally, the whole dataset is fully prepared to be used in various machine learning models.

	primaryid	age	age_cod	sex	wt	wt_cod	prod_ai	route	pt	outc_cod	dechal
0	38853455	65.93566	YR	F	48.00	KG	IMATINIB MESYLATE	Oral	Erythema multiforme	HO	Y
1	39389674	53.00000	YR	M	54.00	KG	CARBOPLATIN	Intravenous drip	Hepatic failure	DE	U
2	39829604	54.00000	YR	M	51.00	KG	CISPLATIN	Intravenous drip	Cachexia	DE	U
3	41148127	57.00000	YR	M	73.00	KG	INFLIXIMAB	Intravenous (not otherwise specified)	Amyotrophic lateral sclerosis	DE	D
4	41155707	15.00000	YR	F	44.45	KG	FENTANYL	Transdermal	Brain injury	DE	Y
5	41546612	23.43600	YR	F	47.63	KG	ETHINYL ESTRADIOL\NORELGESTROMIN	Transdermal	Aortic aneurysm	DE	D
6	57792492	39.00000	YR	F	105.00	KG	INFLIXIMAB	Intravenous (not otherwise specified)	Off label use	HO	Y
7	59585446	61.00000	YR	F	112.15	KG	ETANERCEPT	Subcutaneous	Joint effusion	OT	Y
8	60788732	41.00000	YR	M	85.00	KG	CLONAZEPAM	Oral	Blood cholesterol decreased	OT	U

Figure 3.4 Data Frame

Here we have displayed the whole data frame and all the features we have extracted from the dataset.

	age_cod	sex	wt_cod	prod_ai	route	pt	outc_cod	dechal
count	52464	52464	52464	52190	52464	52464	52464	52464
unique	1	2	1	1930	14	32	6	4
top	YR	F	KG	ADALIMUMAB	Oral	others	HO	Y
freq	52464	27396	52464	1982	30310	35793	35383	22698

Figure 3.5 Data Frame (Descriptive)

In this table we have used python panda library to describe all the values and the attributes we used. Here we got 4 attributes like count, unique, top value, frequency of that value every features.

age	wt	sex_F	sex_M	route_Intramuscular	route_Intraperitoneal	route_Intrathecal	route_Intravenous bolus	route_Intravenous drip	route_Nasal	...	pt_Pyrexia	pt
65.93566	48.00	1	0	0	0	0	0	0	0	...	0	
53.00000	54.00	0	1	0	0	0	0	1	0	...	0	
54.00000	51.00	0	1	0	0	0	0	1	0	...	0	
57.00000	73.00	0	1	0	0	0	0	0	0	...	0	
15.00000	44.45	1	0	0	0	0	0	0	0	...	0	

Figure 3.6 Data Frame (After Get Dummies)

Here we have showed the output of get dummies function. Through this function we have done one hot encoding for every categorical features.

We are using machine learning algorithms to predict the intensity of the ADRs only. The main reason we are using data mining to predict the ADRs itself because machine learning is well suited to predict values with fewer classes, while in our case the dataset contains more than 5000 unique values for the ADR labels. So, using machine learning algorithms to predict 5000 unique classes would bring a very low accuracy value and would not be a good model to use. Instead we used data mining to detect the ADRs and used the detected ADR as an input feature to predict the intensity of ADRs using the machine learning algorithms. As a result, we have created a hybrid model of data mining and machine learning where the output of the data mining method is used as an input to the machine learning model.

We have tried 3 machine learning algorithms which best suit our dataset. They are – Multi-Layer Perceptron, Random Forest, and Support Vector Machine.

3.3.4 Support Vector Machine

Support Vector Machine is a supervised Machine Learning algorithm. Basic concept of this algorithm is finding a hyper plane in order to classify the datasets. There are two kinds of SVM classifier –

a) SVM Linear Classifier

b) SVM non-linear Classifier.

The flowchart of SVM classifier is given below:

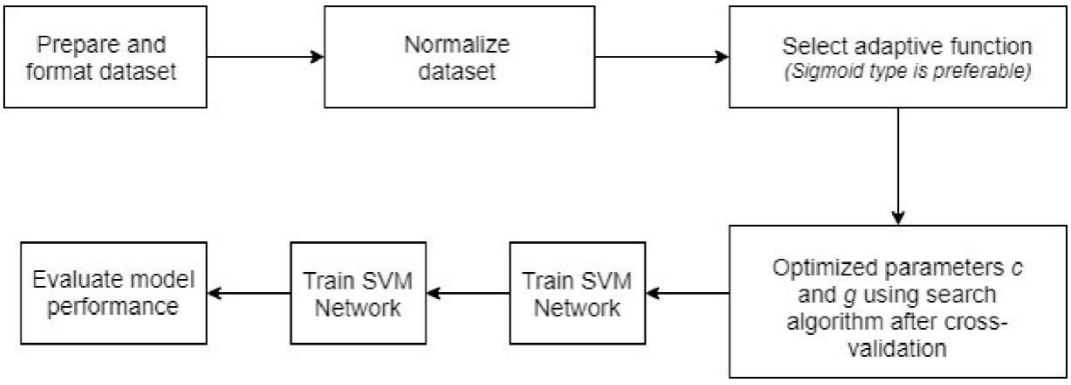

Figure 3.7 Support Vector Machine Work Flow

3.3.5 Random Forest

Random forest is a supervised machine learning algorithm which uses a decision tree approach to create forests and then outputs the maximum occurrence of all the outputs as the final output. The more number of trees the higher chance to get accurate results. RF gives high accurate results and it learns very fast. It is very efficient on large data sets. Random forest uses gini index for deciding the final class of each tree. If data set T contains examples from n classes' gini index, Gini (T) is defined as

$$Gini(T) = 1 - \sum_{j=1}^{n} (P_J)^2 \quad (4)$$

Flow Chart of Random Forest is given below:

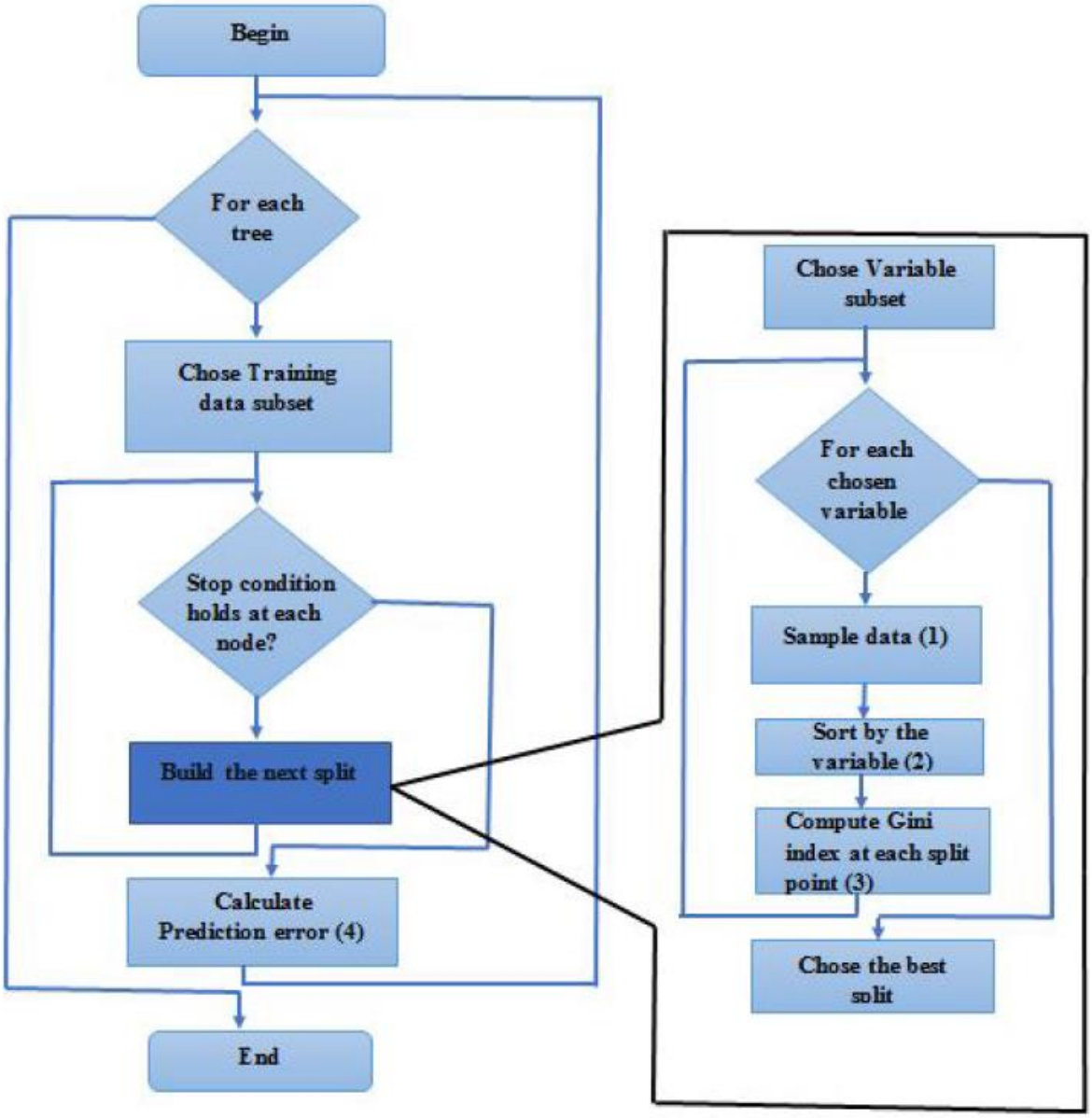

Figure 3.8 Random Forest Flowchart

3.3.6 Neural Network

Multilayer perception is a supervised learning algorithm. MLP has multiple nodes arranged in interconnected layers named input, hidden and output layers. In MLP group of inputs are mapped into asset of desired outputs. The structure of MLP given below

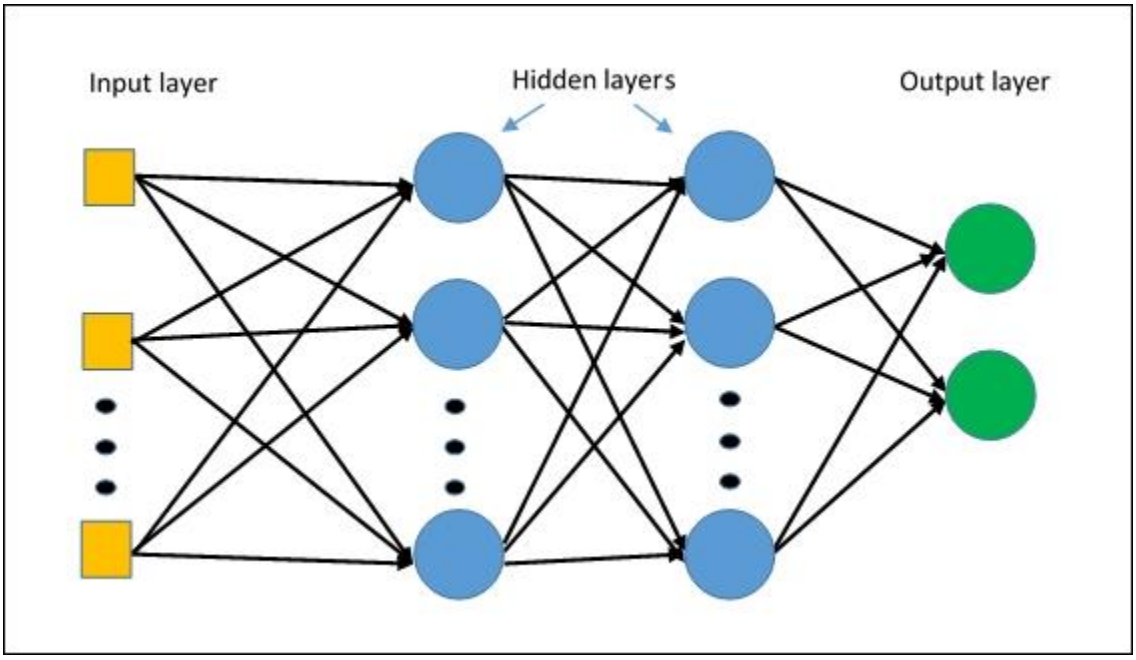

Figure3.9 Multi-Layer Perceptron

The hidden layer cannot be directly accessed. Every layer is made of several neurons. Neurons indifferent layers are connected with those of weight and bias. Equation for the output of neuron 'j' in hidden layer given below,

$$H_j = f(\sum_{i=1}^{n} (w_{ji}x_i + b_i) \quad (5)$$

Here, w_{ji} = weights, b_i = biases, f () = nonlinear activation function

The equation for the network output given below:

$$y = f(\sum_{j=1}^{m} w_{kj}H + b_o) \quad (6)$$

Here, f = output layer neuron activation function, w_{kj}= weight, b_o=bias.

Pseudo code:

Set $Error_{max}$,$Iteration_{max}$,$Rate_{learning}$

Set $NNLayer_{input}$, $NNLayer_{hidden}$, *and* $NNLayer_{output}$

Read $Date_{Training}$, *apply input to the ML-NN network. //ML-NN training phase*

For every input,

Compute the output

Compute error by comparison of acquired value with expected output for the given input

Adjust the weights for all neurons using the obtained error

Repeat the operation until acquired error reached acceptable value, $Error_{max}$

End For loop

End

Example of the System

At first the doctor would enter 2 inputs. A list of symptoms and a list of Drug names. For Example

Table 3.2 Input Example

Drug	Symptoms
Napa	Pain
Taxol	Gastric
Zenol-200	Rash

Then our system would take the list of the drug name and enter it in our data mining model. In the data mining model, the output would be a list of possible ADR. The list would be matched with list of symptoms. Through this a list of ADR will be made as the output of the data mining model.

Then this list is entered in our machine learning model and the intensity of that ADR would be predicted. For example, let's think that the list contains 2 ADR.

- Pain
- Gastric.

Table 3.3 Intensity Example

	Death	Serious Injury	Minor Injury
Pain	0	0	1
Gastric	0	1	0

CHAPTER 4

Result Analysis

4.1 Result of Random Forest Model

Table 4.1 Random Forest Results

Data Management	Random Forest Accuracy
Raw Data	**35%**
Feature Selection	**41%**
Generalization of Adverse Drug Reaction	**45%**
Generalization of Rout(Feature)	**46%**
After removing redundancy	**56%**
Generalizing output into 3 category	**77%**
Generalizing output into 2 category	**91%**

While using Random forest there was a lot of over-fitting of the model. As a Result, there were a huge difference between test set accuracy and training set accuracy. So, to get proper accuracy we used K-fold Cross validation method.

Accuracy when there was redundancy in the data-

Table 4.2 Over fit Results of Random forest

Test Set	Training Set	Cross Validation Set
46%	**89%**	**45.6%**

So, this shows that for working with algorithms like random forest with this kind of data, removing redundancy is must. Or else the model will just memorize the whole data.

4.2 Result of SVM Model

Table 4.3 Results of SVM

Data Management	**SVM Accuracy**
Raw Data	**24%**
Feature Selection	**40%**
Generalization of Adverse Drug Reaction	**40%**
Generalization of Rout(Feature)	**41%**
After removing redundancy	**60%**
Generalizing output into 3 category	**77%**
Generalizing output into 2 category	**91%**

Although, after generalizing the output, the accuracy was almost the same as the Random forest, the initial accuracy was quite low. But the main thing to notice was even with the low accuracy, the model never got over fit and just after removing the redundancy the accuracy increased significantly.

Accuracy when there was redundancy in the data-

Table 4.4 Over fit did not occur in SVM

Test Set	Training Set	Cross Validation Set
46%	**46%**	**46%**

So, this shows that algorithms like SVM has very low chance of overfitting. Even with the redundancy available in the dataset the model never got over fit.

4.2.1 Result of Neural Network

Table 4.5 Results of Neural Network

Data Management	Neural Network Accuracy
Raw Data	**34%**
Feature Selection	**34%**
Generalization of Adverse Drug Reaction	**41%**
Generalization of Route(Feature)	**41%**
After removing redundancy	**51%**
Generalizing output into 3 category	**55%**
Generalizing output into 2 category	**55%**

While starting or research we expected that it but it has the worst of all the algorithm we tested.

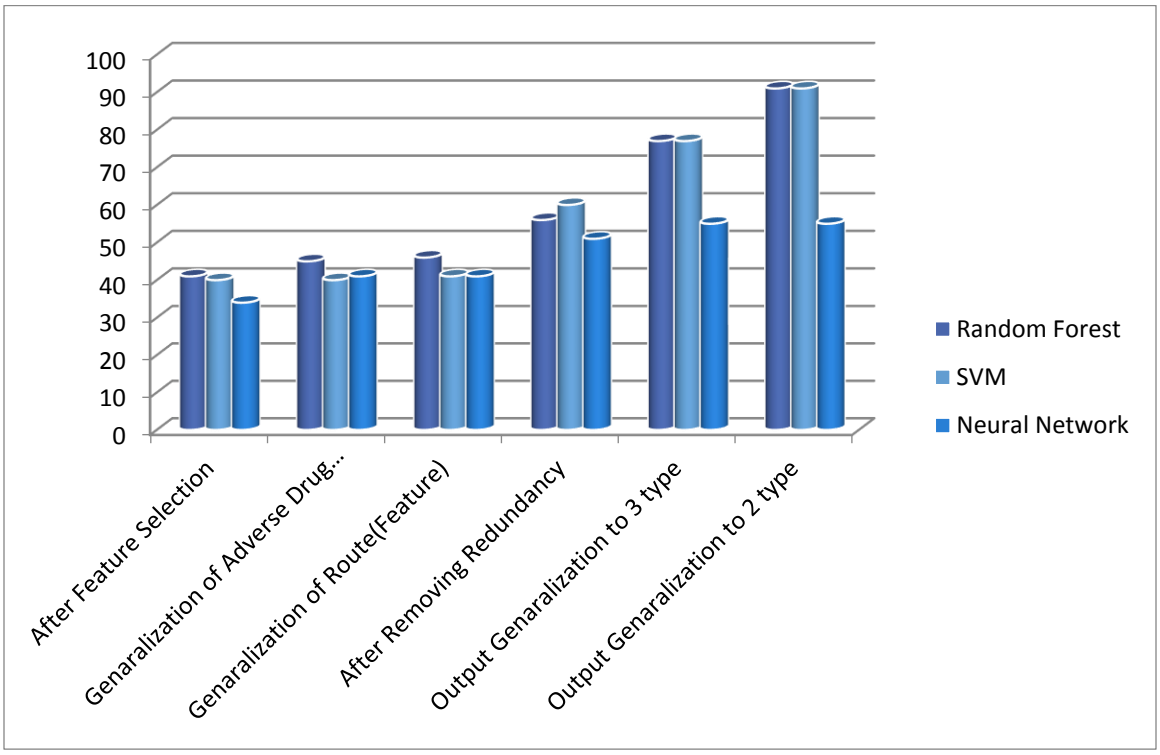

Figure 4.1 Graphical Comparison

4.2.2 Comparison with other related works

Related works have mostly used data mining techniques on their specific datasets to detect ADRs. The usage of the ratios such as Chi-Square and Proportionality Reporting Ratio (PRR), have been commonly used as a data mining technique to detect ADRs. We have used similar ratios in our thesis too. Statistical Analysis is similar for all datasets and it provides results according to that specific dataset. So we were not able to compare with other works from the data mining model. For the machine learning model, we could not find any similar work which used machine learning algorithms for our specific dataset. As a result we could not compare our work accuracy with similar works.

4.2.3 Reason for the initial model had bad results

The main reason why neural network failed to predict the intensity of the ADR is the low interrelationship among the feature.Usually is the best algoritm to use where it features of the model are very highly related with the output.

as mentioned earlier in the report we took a servey from 2 proffesors.Both of them suggested to use features like

- Blood Group
- If the patient is pragnant or not
- If they have any kind of alargy
- Diabetics
- High/low Blood pressure

So we think that if we had feature like these,neural network could have given the highest accuracy for this situations.

CHAPTER 5

Conclusion

Through our work, we have explored various concepts of data mining and machine learning and attempted to come up with a system which will help doctors and pharmacists to perform a safe drug evaluation on a combination of drugs before they prescribe medicine to the patients. We have used Proportionality Reporting Ratio (PRR) and Chi-Square test [10] as our data mining technique to help evaluate the correct combination of safe drugs to be prescribed to the patient and also, we went further ahead with our machine learning concepts to help doctors and pharmacist to be well aware of the outcome of an Adverse Event if it occurs from a combination of drugs. Through this machine learning concepts, using this system, the doctors will be able to avoid accidental deaths from an Adverse Event which might occur from a combination of prescribed drugs and also know the intensity of the adverse event, for example, our system will notify that if an Adverse Reaction might occur for a particular patient, what would be the final condition of the patient – would the patient need to be hospitalized , or would the patient may experience some kind of disability or congenital anomaly or may be the adverse reaction is mild enough to be ignored. Through this hybrid system, the doctors and pharmacists will have a thorough and safe understanding of the combination of drugs and correct the prescriptions accordingly. Point to be noted – a drug is considered safe until an adverse reaction is reported for that drug. But, in the field of medicine there is always scope for uncertainty since each drug can react differently to different specific patients. Our System surely does not give 100% accurate results as not even the doctors are well

capable of that; however, the results that we have obtained is very promising. This system can be used as a compliment tool with the doctor's knowledge and can help aid them in performing a safe drug diagnosis and prescribe the correct combination of medicine to the patient.

5.1 Future Work

The data set itself can be improved in so many ways. A patient's blood group, diabetes information, pregnancy condition (female patients) can be very important factors in determining specific ADRs or to predict the final outcome of an ADR. This information is missing from our dataset. So, if a new dataset can be formed comprising of the valuable information and our above-mentioned concepts are applied to the dataset, a much better result can be achieved. Secondly, since ADRs are person specific mostly and can vary from patient to patient, a ground-breaking result can be achieved if the genetic diagnosis information is available for all the patients and can be used in determining the ADRs. Moreover, if information about the drugs are available on a molecular level, they can be aided with the person's genetic information to create a more stable system for the prediction of Adverse Drug Reaction and their intensity.

REFERENCES

[1] Nebeker, J. R., Barach, P., &Samore, M. H. (2004). Clarifying Adverse Drug Events: A Clinicians Guide to Terminology, Documentation, and Reporting. *Annals of Internal Medicine,140*(10), 795. Doi:10.7326/0003-4819-140-10-200405180-00009

[2] J. M. Beijer, H., & De Blaey, C. (2002). Hospitalizations caused by adverse drug reactions (adr): A meta-analysis of observational studies. *International Journal of Clinical Pharmacy, 24*(2), 46-54. doi:10.1023/A:1015570104121

[3] Wester, K., Jönsson, A. K., Spigset, O., Druid, H., &Hägg, S. (2008). Incidence of fatal adverse drug reactions: A population based study. *British Journal of Clinical Pharmacology,65*(4), 573-579. doi:10.1111/j.1365-2125.2007.03064.x

[4] Huynh, T., He, Y., Willis, A., & Ruger, S. (2016). Adverse Drug Reaction Classification With Deep Neural Networks. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 877-887. Retrieved from http://www.aclweb.org/anthology/C/C16/C16-1084.pdf

[5] Cami, A., Arnold, A., Manzi, S., & Reis, B. (2011). Predicting Adverse Drug Events Using Pharmacological Network Models. *Science Translational Medicine,3*(114), 114-127. doi:10.1126/scitranslmed.3002774

[6] Liu, Y., &Aickelin, U. (2012). Detect adverse drug reactions for drug Pioglitazone. *2012 IEEE 11th International Conference on Signal Processing*. doi:10.1109/icosp.2012.6491898

[7] Zhao, J. (2017). Learning Predictive Models from Electronic Health Records. Retrieved from https://www.diva-portal.org/smash/get/diva2:1067764/FULLTEXT01.pdf

[8] Abin, D., Mahajan, T. C., Bhoj, M. S., Bagde, S., & Rajeswari, K. (2015). Causal Association Mining for Detection of Adverse Drug Reactions. *2015 International Conference on Computing Communication Control and Automation*. doi:10.1109/iccubea.2015.80

[9] Duan, L., Khoshneshin, M., Street, W. N., & Liu, M. (2013). Adverse Drug Effect Detection. *IEEE Journal of Biomedical and Health Informatics,17*(2), 305-311. doi:10.1109/TITB.2012.2227272

[10] Tripathy, A. K., Joshi, N., Kale, H., Durando, M., & Carvalho, L. (2015). Detection of adverse drug events through data mining techniques. *2015 International Conference on Technologies for Sustainable Development (ICTSD)*. doi:10.1109/ictsd.2015.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE:synthetic minority over-sampling technique", Journal of artificial intelligence Research, 2002, pp. 321-357

[12] Monahan BP, Ferguson CL, Cleave ES, Lloyd BK, Troy J, Cantilena LR. Torsade de pointes occurring in association with terfenadineuse. *JAMA* 1990;264:2788–2790.

[13] D. W. Bates, R. S. Evans, H. Murff, P. D. Stetso G. Hripcsak, "Detecting adverse events technology," J. Am. Med. Inf., vol. 10, no Mar./Apr. 2003

[14] Minjoe, S., & Troxell, J. (2017). Preparing Analysis Data Model (ADaM) Data Sets and Related Files for FDA Submission with SAS®. Retrieved from http://support.sas.com/resources/papers/proceedings17/0855-2017.pdf

[15] Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. (Working paper 00/08). Hamilton, New Zealand: University of Waikato, Department of Computer Science

[16] Ren, J. S., & Xu, L. (2015). On Vectorization of Deep Convolutional Neural Networks for Vision Tasks. *29th AAAI Conference on Artificial Intelligence (AAAI-15)*. Retrieved fromhttps://arxiv.org/abs/1501.07338.

[17] Stock, K., Pouchet, L., &Sadayappan, P. (2012). Using machine learning to improve automatic vectorization. *ACM Transactions on Architecture and Code Optimization,8*(4), 1-23. doi:10.1145/2086696.2086729

[18] Evans S et al. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous

[19] Adverse drug reaction reports Pharmacoepidemiol Drug Saf. 2001 Oct-Nov;10(6):483-6, 20th February 2014, 20.00 Hrs

[20] Eva FDA Adverse Drug Event Reporting System, http://www.fda.gov/Drugs/GuidanceCompliancRegulatoryInformation/Surveillance/AdverseDrugEffects/default.html

[21] Emmanuel Chazard, Gr´egoireFicheur, St´ephanieBernonville, Michel Luyckx, and R´egisBeuscart, "Data Mining to Generate Adverse Drug Events Detection Rules", IEEE, VOL. 00, NO. 00, 2011