RESEARCH ARTICLE OPEN ACCESS

ADVERSE MEDICINE REACTION ANALYSIS USING SUPPORT VECTOR MACHINE MODEL

C.Mani MCA., M.E.*, V.Madhan Kumar**
* AssociateProfessor, Department of CSE, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.
Email: cmanimca@gmail.com
** Final MCA, Department of MCA, Nandha Engineering College (Autonomous), Erode, Tamilnadu, India.
Email: vmadhan1999@gmail.com

****

Abstract:

Adverse Drug Reaction (ADR) is one in every of the various uncertainties that are considered a fatal threat to the pharmacy industry and therefore the field of diagnosing. Utmost care is taken to check a brand new drug thoroughly before it's introduced and made available to the general public. However, these pre-clinical trials don't seem to be enough on their own to confirm safety. The increasing concern to the ADRs has motivated the event of statistical, data processing and machine learning methods to detect the Adverse Drug Reactions. With the provision of Electronic Health Records (EHRs), it's become possible to detect ADRs with the mentioned technologies. during this work, we've proposed a hybrid model of information mining and machine learning to spot different Adverse Reactions and predict the intensity of the end result. we've used the Proportionality Reporting Ratio (PRR) together with the precision point estimator test called the Chi-Square test to search out out the various relationships between drug and symptoms called the drug-ADR association. This output from the information mining technique is employed as an input to the machine learning algorithms like Random Forest and Support Vector Machine (SVM) to predict the intensity of the end result of ADR, looking on a patient's demographic data like gender, weight, age, etc. during this work, we've achieved an accuracy of 91% to predict 'death' because the outcome from an ADR.

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

****

I. INTRODUCTION

In the planet of bioscience and medicines, Adverse Drug Reaction (ADR) has always been a crucial field of research. Adverse drug reaction means the injury from the utilization of a drug. These injuries can extend from minor injuries like rash to major life-threatening reactions. Confusion occurs mostly between ADR and Side Effect where ADR is that the reaction caused by the drug used at normal doses for particular symptoms. Point to be noted: a wrongful overdose of medicine isn't considered as an ADR case.

Every Year quite 200 thousand deaths are reported due to ADRs. Though ADR are often identified very easily after the occurrence, predicting ADR has always been a large challenge for researchers. Worldwide, around 4.9% of hospital admissions are the results of ADRs and this number is as high as 41.3% in some areas. In Sweden, ADRs are the seventh most typical explanation for death. While drugs are thoroughly tested clinically before they're released on the

market, many unknown side effects are discovered after they need been used over time by various patients. Moreover, as ADR varies from person to person, predicting ADR can sometimes be as hard as impossible even for doctors.

Nowadays almost 73% of the people round the world take different medications. Among these, almost 29% of those medicines have different varieties of adverse drug reactions. FDA statistics show that nearly 7000 of those ADRs have caused death in recent years. Through this paper, we tried to predict the ADR and its intensity so necessary precautions are often taken before prescribing any sorts of medications.

In recent years, a large amount of Electronic Medical Records (EMR) are available on several platforms. As a result, various methods of information science is implemented for the detection of ADR. Various methods of machine learning and data processing have already been implemented to predict the possible ADR from a particular drug. However, during this paper, we've proposed a hybrid model of machine learning and data processing to predict the ADR and its severity. We contributed totally on the machine learning part where we took the detected ADR from the already existing, statistical data processing techniques and used it to seek out the severity of the reaction on various unique patients.

II. LITERATURE REVIEW

The aforementioned challenges motivated a series of works that apply data processing and machine learning approaches to return up with various solutions using different datasets to suit per the researcher's needs. Google recently worked on a Twitter dataset to detect ADEs from posts on Twitter employing a merged sort of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). They used a binary classifier to represent the end result. Another researcher proposed a Predictive Pharmacosafety Networks (PPNs) to detect likely unknown ADE. He used the prevailing drug safety information from a well known data set of drug safety in 2005 to coach a logistic regression model to detect unknown ADEs. In 2012, a probe used the 'THIN' database to form a model using the feature matrix and have selection to spot ADRs for a selected drug called "Pravastatin".

A research in 2017 proposed a model using Electronic health records and train them within the Random Forest algorithm to predict future unknown adverse drug reactions. The random forest algorithm uses a call tree approach to make a forest and choose the end result with the utmost count as their output. a knowledge mining technique called "Casual Association Rule" has been accustomed find a cause-effect relationship between the drug-symptom pair which will be accustomed detect ADRs. In 2013, another researcher named Duan proposed two novel models - a likelihood ratio model and a Bayesian network model - to form a pattern discovery method to spot adverse drug reactions .

A new research used data processing techniques to seek out a relationship between a mixture of medication and their possible ADRs using the Chi-Square and Proportionality Reporting Ratio (PRR) as their basis to search out the link. This paper depends heavily on this idea as we've got used similar data processing techniques for our work. In most of the researches that we've found, detecting the ADRs was the most concern, but because of the vastness and uniqueness of the ADRs, it's almost impossible to detect it with a promising precision. This can be the most reason why we've proposed the aforementioned hybrid model.

III. RESEARCH METHODOLOGY

A. Preprocessing & Data Description

The model proposed during this paper used the publicly available dataset from the Food and Drug Administration (FDA) . For the data-mining a part of the model, we've used the "drugname" feature from the "Drug" table and "pt" from the "Reaction" table. the link between the attributes is shown in Fig. 1.

For the machine learning a part of the model we took into consideration all the tables and performed various feature selection algorithms like Minimum-redundancy maximum-relevance (mRMR), Pearson's Correlation and python's built-in panda library called "Feature Selection", to search out out

the simplest features which are most relevant to the end result class. Moreover, we consulted experts within the field of life science and prioritized the features they suggested as most relevant to the end result feature.

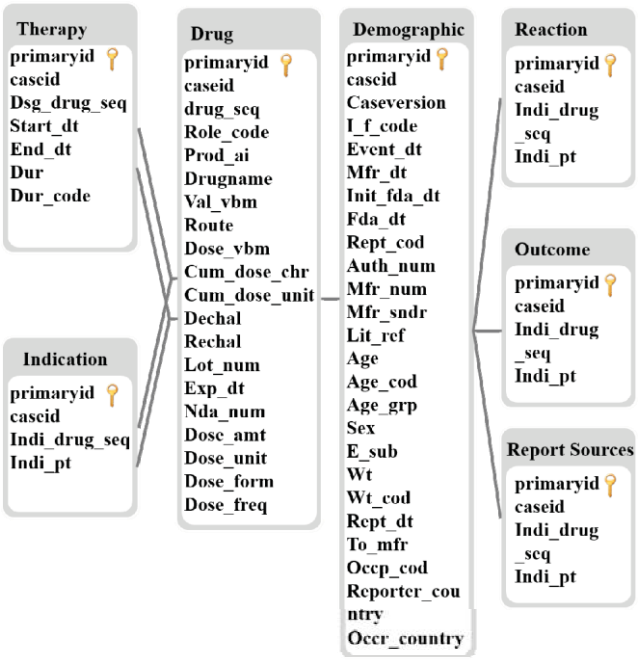

Fig. 1. Dataset Tables

Finally, after combining the results from all the feature selection methods, we selected 6 features which are most relevant to the end result class and used them as training features in various machine learning algorithms.

Most supervised machine learning algorithms absorb numerical values for the training features. Methods like One-Hot Encoding, Vectorization and Label Encoding are wont to convert the explicit columns within the training feature set into numerical values whenever required.

Training Features:

- "Age" – Patient's numerical value for age at the time of event occurrence.
- "Sex" - Patients' sex (Male and Female).
- "Wt" - the Numerical value of patient's weight.
- "Route" -The route of the drug intake within the patients' body
- "Dechal" - Indicates whether the adverse reaction subsided when drug in-take was terminated
- "Pt" - "Preferred Term". it's the medical terminology for the adverse drug reaction.

Outcome Tags

- "Death" – The patient will die from the ADR if occurs.
- "Major Injury" – The patient might must be hospitalized because of life-threatening situations.
- "Minor Injury" – The patient might experience some style of abnormal behaviour within the body like pain or swollen skin or rash, etc.

B. Proposed HybridModel

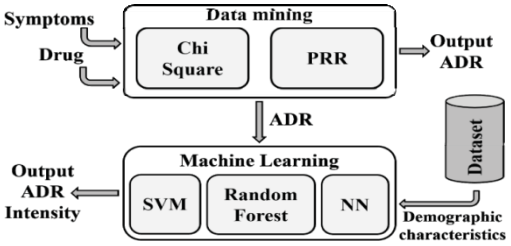

Fig.2.HybridModel

Fig. 2 shows a mixture of information mining techniques and machine learning algorithms to make a hybrid model for identifying the ADRs and predict their severity.

C. data processing

To find an association between the drug and therefore the symptoms in control of the prescribed drug, data processing association techniques like Chi-Square and
Proportionality Reporting Ratio (PRR) are used which give numerical values for the model. If the values exceed a particular threshold, that specific symptom is identified as an ADR for a specific drug.

TABLE I. RATIO VARIABLE CALCULATION

Category	Specific Reaction	AllOther Reactions	Total

Specific Drug	A	B	A+B
All Other Drugs	C	D	C+D
Total	A+ C	B+ D	N=A+B+ C+D

The general criteria to run the PRR are as follows:

- Value A refers to the quantity of occurrences, where a particular drug P, causes a selected Adverse Reaction, R.
- Value B refers to the quantity of occurrences, where specific drug P causes the other Adverse Reactions but R.
- Value C refers to the amount of occurrences, where the Adverse Reaction R is caused by any other drug but P.
- Value D refers to the amount of occurrences, where the other Adverse Reactions but R is caused by the other drugs but P

PRR= (A / (A+B)) / (C / (C+D)) (1)

Chi Square X2 = (AD-BC)2+(A+B+C+D)/ [(A+B) (B+C) (C+D) (A+D)] (2)

The threshold value for Chi-Square and PRR:

- PRR > 2
- Chi-Square >4

TABLE II. ADRDETECTION

Drug Name	Preferred term	Chi Square	PRR
SALINE	Musculoskeletal Stiffness	1.2758	0.00
DECADRON	Musculoskeletal Stiffness	1.2758	0.00
GASTER	Musculoskeletal Stiffness	1.2758	0.00
TAXOL	Musculoskeletal Stiffness	1.2758	0.00
DIPHENHYDRAMINE HCL	Musculoskeletal Stiffness	1.2758	0.00
ABILIFY	Anxiety	20.5213	3.5794
CLOZAPINE	Agitation	1.2314	0.9126
VALPROIC ACID	Agitation	1.2314	0.9126
ATENOLOL	Agitation	1.2314	0.9126

The drug reactions with values greater than the brink are matched with already given symptoms by the user. If thereis any cross match, then that's considered as a possible ADR.

A.Machine Learning

The identified ADR from the info mining model is employed in concert of the input features to coach the machine learning algorithms like Random Forest (RF), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP), with the aim of predicting the severity of the previously identified ADR. As explained within the Processing and Data Description part, we came up with 6 input features and merged them together into a dataset to coach the machine learning models. Moreover, the merged dataset needed to be processed further to create it suitable for training. We performed "Generalization" on a number of the training features like "Pt" and "Route" and reduced the "Redundancy" of the dataset.

Generalization

The feature called "Pt" could be a categorical column with quite 5000 unique names. to form it suitable as a training feature we reduced the quantity of unique values, as that they had to be converted to numerical values within the later stage. We generalized the unique values during this input feature into subgroups. for instance, the "Pt" column had unique names for various kinds of "Anaemia" like "Iron Deficiency Anaemia", "Haemorrhagic Anaemia", "Aplastic Anaemia"and many others; we generalized of these unique names into just "Anaemia". Moreover, the "Route" column had over 62 unique names, of which, some only occurred once or twice in the whole dataset. To prevent any outliers, we removed the least occurring ones and selected 16 most occurring names.

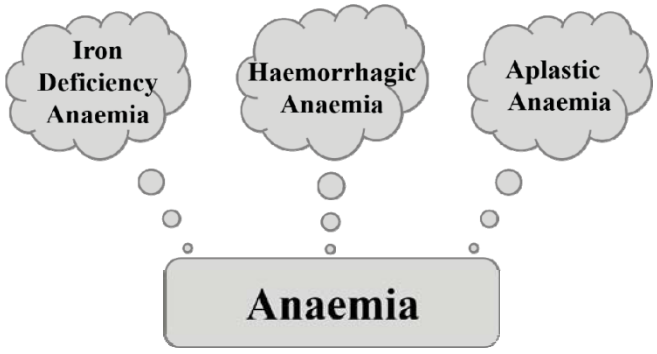

Fig. 3. Generalization

Redundancy

A lot of the records within the dataset were overlapping. for instance, a patient had an ADR occurrence and was hospitalized. The dataset encompasses a record of that patient within the serious injury class. However, later the patient died, and therefore the dataset also contains a record of the identical patient within the "death" class.

Therefore, to scale back any anomaly within the results we removed the previous records of the identical patient and kept the concluding record of that patient. this is often how we reduced redundancy from the full dataset.

Finally, the merged dataset with 6 input features is processed and prepared to be fed into the machine learning algorithms for the training purpose.

Model Selection For Training

Since we are solving a classification type problem where we are predicting 3 different output classes, we'd like classification algorithms to make a training model.

Moreover, this dataset falls under the supervised machine- learning category, where, we've got input variable(X), output variable (y) and so we use an algorithm to be told the mapping function from input and output.

Random Forest, Support Vector Machine and Multi-Layer Perceptron are few of the foremost efficient machine learning algorithms that fulfill our solution criteria. We used these models to coach our dataset and evaluated how each performed on the task of prediction.

A.Proposed SystemImplementation

At first, the doctor would enter 2 inputs. an inventory once or twice within the whole dataset. to forestall any outliers, we removed the smallest amount occurring ones and selected 16 most occurring names.

of symptoms and a listing of drug names prescribed for those symptoms. For Example:

TABLE III. LIST OF SYMPTOMS ANDDRUGS

Drug	Symptoms
Atenolol	Asthma
Metoprolol	Scar
Taxol	Gastric
Zenol-200	Rash

Our system would take the list of the drug names and enter it in our data-mining model. The data-mining model would offer a listing of all possible ADRs associated with those drug names. That list is then cross-matched with the symptoms input. If there's any cross-match, the case is identified as an adverse reaction case, or if not matched, it's considered a secure case.

TABLE IV. LIST OF ADVERSE REACTIONS FOR DRUGATENOLOL

Reaction	PRR	Chi-Square
Insomnia	2.427	7.231
Asthma	3.052	5.131
Skin Discoloration	5.188	9.905
Scar	4.612	8.195
Emotional Disorder	3.861	6.011

TABLEV. LIST OF ADVERSE REACTIONS FOR DRUGMETOPROLOL

Reaction	PRR	Chi-Square
Deformity	4.696	17.502
Rash	2.082	4.204

Joint Stiffness	5.366	10.579
Cardiomegaly	2.981	4.941
Emotional Disorder	3.744	9.856

As shown in Table IV and Table V, from the list of ADRs associated with the Drugs "Atenolol" and "Metoprolol" 3 ADRs matched with the input Symptoms. Therefore, this can be a case of an adverse event. For this adverse event case, the identified ADRs are now entered in our machine-learning model as an input feature and also the severity of these ADRs are predicted.

- Asthma
- Scar
- Rash

TABLEVI.THE INTENSITY OF ADRDETECTION

Symptoms	Death	Serious Injury	Minor Injury
Asthma	0	1	0
Scar	0	1	0
Rash	0	0	1

V. RESULT ANALYSIS

After calculating the worth of Chi-Square and PRR for all the drugs within the dataset, as shown in Fig. 4, we've got found that only 821 drugs don't have any record of ADRs within the dataset. So, these are considered safe for now until an ADR is reported against these drugs.

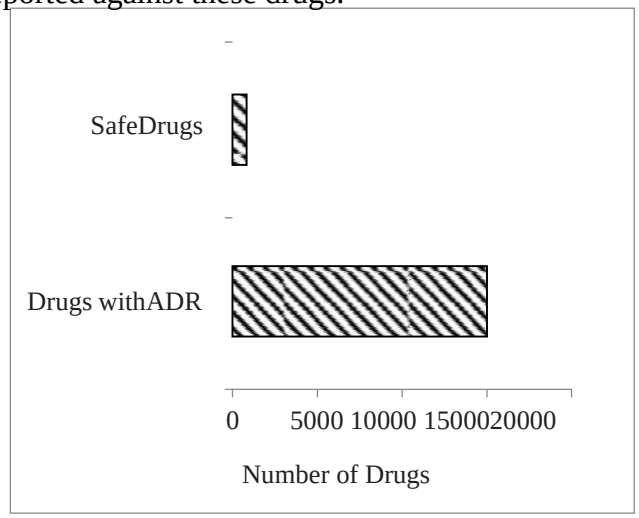

Fig. 4. Number of drugs in the dataset

More than 10000 drugs have adverse reaction records within the dataset, but that doesn't mean that the drugs can not be considered safe during a combination of prescribed medicines. As long as there in no cross-match between the ADR list of the drug and therefore the patient's symptoms, the drugs are safe to be prescribed to the patients. The intensity of the ADR cases is evaluated through machine learning.

We have used 10-fold cross validation method to coach our machine learning models in order that we will avoid overfitting as there's lots of outlier values in our dataset. The dataset was split into training and test a part of 80 attempt to 20 % respectively. As shown in Fig. 5, after splitting into training and test set we got 176 thousand records as our test set. However, after training our model with the training set, we used our test set for verification. As shown in TABLE VII, Random Forest and SVM could predict death with 91% accuracy, minor injury and major injury with 81% accuracy. the particular number of occurrences for death, major and minor injury was 31%, 51% and 46% of the full dataset, respectively, as shown in Fig. 5.

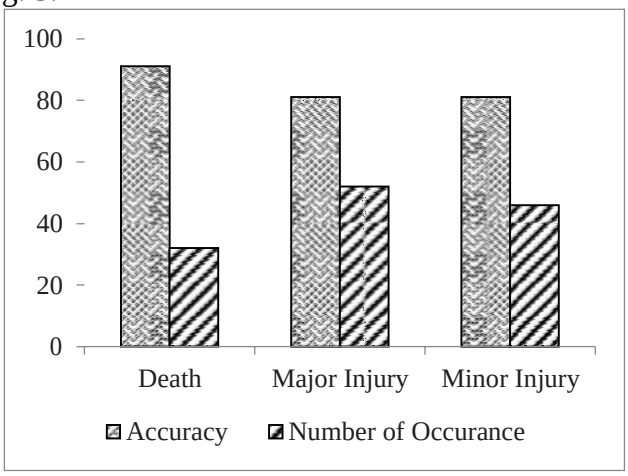

Fig. 5. Number of occurences of different ADR intensity

Both the Random Forest (RF) and Support Vector Machine (SVM) performed quite well but we had all-time low accuracy within the Multi-Layer

Perceptron (MLP) model for detecting both severity and death.

For better results, the dataset itself is improved in such a large amount of ways. A patient's blood type, diabetes information, pregnancy condition (female patients) are often important factors in determining specific ADRs or to predict the result of an ADR. This information is missing from our dataset. So, if a replacement dataset will be formed comprising of the precious information and our above-mentioned concepts are applied to the dataset, a way better result is achieved. Secondly, since ADRs are person specific mostly and might vary from patient to patient, a groundbreaking result are often achieved if the genetic diagnosis information is on the market for all the patients and may be employed in determining the ADRs. Moreover, if information about the drugs is offered on a molecular level, they will be aided with the person's genetic information to form a more stable system for the prediction of Adverse Drug Reaction and their intensity.

TABLEVII. THE ACCURACY OF DIFFERENTMODELS

Category	Random Forest	Support Vector Machine	Multi-Layer Perceptron
Death Prediction Accuracy	91.4 %	91.3 %	55.7 %
Intensity prediction Accuracy	81.7 %	81.6 %	53.6 %

CONCLUSION

Through this paper, we've explored various concepts of knowledge mining and machine learning and attempted to return up with a hybrid model which will help doctors and pharmacists to perform a secure drug evaluation on a mixture of medicine before they prescribe medicine to the patients. we've used the Proportionality Reporting Ratio (PRR) and Chi-Square test as our data processing

technique to assist evaluate the proper combination of safe drugs to be prescribed to the patient and also, we went further ahead with our machine learning concepts to assist doctors and pharmacist to be tuned in to the result of an Adverse Event if it were to occur from a mix of medication. A drug is taken into account safe until an adverse reaction is reported for that drug. However, within the field of medication, there's always scope for uncertainty since each drug can react differently to different specific patients. Our System surely doesn't give precise results, as not even the doctors are well capable of that; however, the results that we've obtained are very promising. this technique is used as a complimentary tool with the doctor's knowledge and may help aid them in performing a secure drug diagnosis and prescribe the right combination of medication to the patient.

REFERENCS

[1] J. Nebeker, P. Barach and M. Samore, "Clarifying Adverse Drug Events: A Clinician's Guide to Terminology, Documentation, and Reporting", Annals of medical specialty, vol. 140, no. 10, p. 795, 2004.

[2] H. Beijer and C. de Blaey, Pharmacy World and Science, vol. 24, no. 2, pp. 46-54, 2002.

[3] K. Wester, A. Jönsson, O. Spigset, H. Druid and S. Hägg, "Incidence of fatal adverse drug reactions: a population based study", British Journal of Clinical Pharmacology, vol. 65, no. 4, pp. 573-579, 2008.

[4] T. Huynh, Y. He, A. Willis and S. R¨uger, "Adverse Drug Reaction Classification With Deep Neural Networks", in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 2016, pp. 877-887.

[5] A. Cami, A. Arnold, S. Manzi and B. Reis, "Predicting Adverse Drug Events Using

Pharmacological Network Models", Science Translational Medicine, vol. 3, no. 114, pp. 114ra127-114ra127, 2011.

[6] Y. Liu and U. Aickelin, "Detect adverse drug reactions for drug Pioglitazone", 2012 IEEE 11th International Conference on Signal Processing, 2012.

[7] J. Zhao, "Learning Predictive Models from Electronic Health Records", Ph.D, Stockholm University, 2017.

[8] D. Abin, T. Mahajan, M. Bhoj, S. Bagde and K. Rajeswari, "Causal Association Mining for Detection of Adverse Drug Reactions", 2015 International Conference on Computing Communication Control and Automation, 2015.

[9] L. Duan, M. Khoshneshin, W. Street and M. Liu, "Adverse Drug Effect Detection", IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 2, pp. 305-311, 2013.

[10] A. Tripathy, N. Joshi, H. Kale, M. Durando and L. Carvalho, "Detection of adverse drug events through data processing techniques", in 2015 International Conference on Technologies for Sustainable Development (ICTSD), Mumbai, India, 2015, pp. 01-06.

[11] D. Mathews, "Torsades de Pointes Occurring in Association With Terfenadine Use", JAMA: The Journal of the American Medical Association, vol. 266, no. 17, p. 2375, 1991.

[12] S. Minjoe and J. Troxell, " Preparing Analysis Data Model (ADaM) Data Sets and Related Files for FDA Submission with SAS®", 2017.

[13] M. A. Hall, " Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning", in ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, city, CA, USA, 2000, pp. 359-366.

[14] R. Jimmy and X. Li, "On Vectorization of Deep Convolutional Neural Networks for Vision Tasks", in 29th AAAI Conference on computing (AAAI-15), Texas, USA, 2015, pp. 25-30.

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

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