iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36405

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Disease Analysis Using Medicines' Sales Data

Mr. Omkar S. Mohite[1], Dr. Surekha Kohle[2]
[1, 2, 3]*Master of Computer Application, Department Veermata Jijabai Technological Institute (VJTI Mumbai) University of Mumbai, Maharashtra, India*

Abstract: The medical industry is one of the vast organized industries, which deals with various kinds of health-related products and services. Additionally, this industry also contributes in Research and Development of Medicines to cure diseases. Hence, they also have their sales data for a particular area, using some kind of aggregation, predication and clustering we can use the sales, association data for finding the outbreak of a daisies or we can check the sales of that particular area and demand for a particular medicine. Using this data, we can make some precautionary measures on that area plus we can analyze the effectiveness of the particular medicine. Plus, this data helps in the Research and Development of new medicine.
The medical industry is one of the vast organized industries, which deals with various kinds of health-related products and services. Additionally, this industry also contributes in Research and Development of Medicines to cure diseases. One of the important aspects of every organization is the Sales, this industry too has different sales data for every particular area. With the help of aggregation, prediction and clusters we can use the sales data associated with medicines to find outbreak of any disease or we can verify sales of any particular area and look at the demand for a particular medicine. Using this data, we can take early precautionary measures in those areas, also we can analyze the effectiveness of that particular medicine. This new kind of data will be beneficial for Research and Development of new medicines.

I. INTRODUCTION

A. *About The Subject*

In the medical industry various kinds of Research and Development takes place. This market is very Price and Time sensitive, where if any new disease is spread then to create a cure for that disease is a very tedious task. This research paper suggests an approach where the supply chain for providing the medicine can be simplified in those infected areas.

B. *About The Disease*

A disease is any kind of infection inside a human body in the form of any bacteria or other virus. This infection could have occurred due to any reaction in a human body causing change in body temperature, unwillingness to eat food, lack of taste, etc. Every infection has different symptoms, this arises as the immune system of the human body tends to be weak.

C. *Supply Chain Of The Medicine*

The supply of any medicine is a very difficult task in large populated countries like China, India and others. As the scale for supply of the medicine is very large and being very time sensitive market, the supply of medicine has to be done in an efficient way. Supply starts with the plant where the medicine is produced and then it is sent to the citizens of country



Fig 1. Supply Chain of Medicine [3]

The medical supply chain is such organized that a barcode is made to find and map the route of the medicine where it is supposed to reach efficiently.

D. *Relation Between Medicine And Disease*

A Medicine is a cure over a disease made from chemical compounds which helps to increase immunity and prevents the disease. The relation between a disease and medicine is such where a single medicine could be used to cure one or many diseases, whereas only few medicines could be used to cure certain types of diseases. This medicine reacts in such a way that the contents in it help to produce more immune cells reacts inside a human body to fight with the disease.

©IJRASET: All Rights are Reserved

E. About The Project

In today's scenario, the pharmaceutical companies analyze the data provided by distributors and hospitals based on their sales figures. This approach by the companies helps them to map their profit and demand for future strategies, which can have a different angle to analyze the data from government's perspective for betterment of citizen's health. The research paper suggests the pharmaceutical companies to consider a different approach. The Government would analyze the data which will assist to state predict which part of the state suffers through any particular disease based on the sales figure of the medicines sold in those areas.

F. Data Association

The objective of Data Association is to verify the relation between medicines with diseases. The research paper proposes an approach to analyze data which would help to show which area suffers through any disease by observing the number of medicines sold in those areas. This approach could be followed as- When raw data from distributors and hospitals of particular areas would be collected by the Government, then they shall interprete the data using the sales figures of medicines in those areas and analyze the data in such a way that would depict which disease is spread over that particular area.

II. LITERATURE REVIEW

A. Drugs analysis for Outbreak Detection

1) *Introduction:* The aim of the study was to summarize evidence for the added value of drug sales data analysis for the surveillance of infectious diseases.
2) *Strengths:* Drug sales data analysis may overcome the limitation of poor specificity when groups of drugs are exclusively used for the disease or disease syndrome of interest. Furthermore, drug sales data may earlier capture changing population health status, as over-the-counter (OTC) sales and a dense network of pharmacies in most developed countries make drugs easily accessible to patients at the earliest appearance of their symptom. Most of the studies focused on respiratory illnesses (17 studies) or gastrointestinal illnesses (11 studies). Only two other studies evaluated surveillance of pertussis and syphilis. Three types of studies were defined: retrospective descriptive studies, drug selection studies and prediction studies Nineteen of the 27 studies were descriptive retrospective studies assessing the strength of the correlation between drug sales and reference surveillance data of the corresponding disease or evaluating outbreak-detection performance. Five studies used statistical algorithms to select groups of drugs that were closely associated with clinical surveillance data of a given disease and that would be most appropriate for future drug-sales-based surveillance. In a third group of three studies, the authors developed and evaluated statistical models to predict clinical surveillance data based on drug sales data.
3) *Location of the Sample:* All the studies were conducted in developed countries or area. Most of the studies were set in the United States (n = 16 studies, 59%), followed by Canada (n = 4), France (n = 3), Japan (n = 3), the Netherlands (n = 1) and England (n = 1). Only one study was conducted in more than one country[1]

Methodology and results of drug selection studies included in a literature review of drug sales data analysis for surveillance of infectious diseases

Author	Disease	Method	Results of the algorithm evaluation
Pelat et al. [33] 2010	GI	Hierarchical clustering procedure ,CUSUM	Identification of 4 therapeutic classes relevant to gastroenteritis outbreak detection. Detection performance of a multiple voter algorithm: sensibility 100%, specificity 95%, timeliness 1.7 weeks.
Cami et al. [31] 2009	ILI	Aggregate mining algorithm	Identification of product categories with outbreak detection performance superior to predefined categories and more strongly correlated with the disease data.
Wallstrom et al. [34] 2007	ILI	Unsupervised time-series clustering algorithm	Distinction between OTC products for allergy and OTC products for influenza symptoms
Li et al. [32] 2005	ILI/GI	Canonical correlation analysis	Identification of eight diagnoses that have strong association with electrolyte sales (r = 0.96)
Magruder et al. [21] 2004	ILI	Unsupervised stepwise clustering algorithm	Identification of 16 OTC product groups with similar historical trends

Abbreviations: GI Gastrointestinal, *CUSUM* Cumulative sum control chart, *ILI* Influenza-like illness, *OTC* Over-the-counter drugs.

Fig. 2. Methodology and results of drug selection [1]

©IJRASET: All Rights are Reserved

4) *Outcome:* The evidence gathered in this systematic literature review suggests that drug sales data analysis can be a useful tool for surveillance of acute respiratory and gastrointestinal infections. As could be expected, prescribed drug sales data were strongly correlated with clinical case reporting. No lead time was observed, which is consistent with the fact that patients purchase drugs after seeing a healthcare professional. Analysis of prescribed drug sales data may nevertheless have an additional utility for epidemic detection, as these data might be available with a shorter delay than clinical surveillance data
5) *Benefits of the Study:* By improving the timeliness of epidemic detection compared to clinical data and giving information from a larger part of the population, drug sales data can be an additional source of information for already monitored diseases. Besides, drug sales data analysis could have its greatest value in the surveillance of diseases for which clinical surveillance is cumbersome and costly, or where substantial under-reporting is suspected. To confirm the selected drug group as a valid proxy of disease, clinical surveillance may be conducted for a defined period in a representative population. Examples of diseases for which this would be useful are typically varicella, urinary infections, allergies/asthma, and parasitic diseases.
6) *Limitations:* Ideally, the drugs to be monitored should be specific to the disease and widely used to treat it in order to maximize the sensitivity of the signal. For example, benzylpenicillin benzathine 2.4 MUI is the quasi-exclusive treatment for syphilis infection [1] and is a good candidate. In contrast, the treatment of measles is mostly symptomatic without a specific drug, which makes this disease unattractive for this approach. Another limitation applies to diseases that are usually treated in hospitals or specialized centers, such as tuberculosis. Surveillance based on drug sales, may not be appropriate to accurately estimate incidence of diseases, as the source population size is not precisely known. Moreover, it may be difficult to link the number of drug packages sold to the number of patients with disease. However, the method is very efficient to determine temporal dynamics of a situation and to detect abnormal phenomena. Surveillance based on drug sales is therefore well adapted to diseases with seasonal variations such as norovirus gastroenteritis, influenza and other infectious respiratory agents, or community outbreaks (foodborne illnesses, waterborne illnesses, hepatitis A, etc.). [1]. Drug sales can be influenced by store promotions, sales period (holidays, weekends), and the media. Also, we do not know whether people buy medications to treat a disease they currently have or a disease they fear they may have in the near future. For example, during the media coverage of avian influenza A (H5N1) in the US, an increase in antiviral medications sales was observed [40], which corresponded to stockpiling behavior of the population.
7) *Scope:* By improving the timeliness of epidemic detection compared to clinical data and giving information from a larger part of the population, drug sales data can be an additional source of information for already monitored diseases. Besides, drug sales data analysis could have its greatest value in the surveillance of diseases for which clinical surveillance is cumbersome and costly, or where substantial under-reporting is suspected. To confirm the selected drug group as a valid proxy of disease, clinical surveillance may be conducted for a defined period in a representative population. Examples of diseases for which this would be useful are typically varicella, urinary infections, allergies/asthma, and parasitic diseases.
8) *Conclusion:* This review suggests that the analysis of drug sales data is a promising method for surveillance and outbreak detection of infectious diseases. It can possibly trigger a flare-up alarm than most surveillance systems. However, the main challenges consist in proper choice of pointer drug gatherings and the approval of this methodology for sicknesses for which no or poor quality clinical observation information exists. The handiness of the methodology additionally relies upon the accessible assets and the association of the medical care framework. Drug sales databases with real-time or near real-time data transmission are available in several countries; future studies should be encouraged to expand their use on other infectious diseases.

B. *Drug Disease Association*

1) *Need:* Building up another medication is an unpredictable, tedious, and costly cycle, which ordinarily continues through fundamental compound testing, pre-clinical and creature tests, clinical exploration, and Food and Drug Administration (FDA) survey, before it at long last yields another medication that arrives at the market following 10–15 years, costing around 0.8–1.5 billion dollars. Indeed, even with a generous time responsibility and capital venture, the effective advancement of another medication is as yet connected with extensive dangers. Since the quantity of new medications affirmed by the FDA has been declining since the 1990s, there is an urgent need to find alternative approaches that will reduce the development costs. Drug repositioning refers to the identification of new indications for drugs that have been approved by regulatory agencies. Compared to the development of a new drug for a certain indication, drug repositioning can shorten the drug development cycle to 6.5 years at the cost of approximately 0.3 billion dollars due to the known safety, tolerability, and efficacy profile of the drug candidate. Identifying new indications for existing drugs may reduce costs and expedites drug development. Drug-related

©IJRASET: All Rights are Reserved

disease predictions typically combined heterogeneous drug-related and disease-related data to derive the associations between drugs and diseases, while as of late created approaches coordinate numerous sorts of medication highlights, yet neglect to consider the variety suggested by these highlights.

2) *Introduced Model:* In this study, a model was presented with a new method, known as DivePred-for predicting potential drug–disease associations. DivePred profoundly coordinates not just the projection of numerous medication highlights in low-dimensional space yet in addition the variety of medication highlights. Anticipating numerous high-dimensional medication highlights into a similar measurement as the illness helps with estimating the distance between the medications and the sicknesses, which is a critical parameter for the possibility of a drug–disease association. The chemical substructures of the drugs, the target protein domains, and the ontology annotation of the target gene, alongside its related infection explanations mirror the qualities of the medications from alternate points of view. Therefore, retaining the diversity of multiple drug features could fully integrate information from different drug views.
3) *Model Differentiation (Speciality):* The created model was a unified model and it had the ability to develop an iterative optimization algorithm to derive drug–disease association scores. Experimental results based on cross-validation indicated that DivePred achieved better prediction performance than several state-of-the-art methods.
4) *Evaluation Metrics:* A five-fold cross-validation to evaluate the performance of DivePred in predicting potential drug–disease associations. The known drug–disease associations were randomly divided into five equal subsets, four of which were used to train this model, while the remaining set was used to perform the test. In each cross-validation, X(4) contained only the drug–disease associations of the training set, and R4 was calculated based on the known associations in matrix X(4). For a certain drug ri($1 \le ri \le Nr$), its associated diseases in the test set was called the positive sample, and the other unmarked diseases were called negative samples. In the test results, a high positive sample rate of drug was correlated with an improved predictive performance for this drug. A threshold θ was set, and when the score obtained by the sample estimate was higher than θ, it was identified as a positive example; otherwise, it was identified as a negative example. The TPRs (true-positive rates) [2] and the FPRs (false-positive rates) [2] under various θ could be calculated as follows,

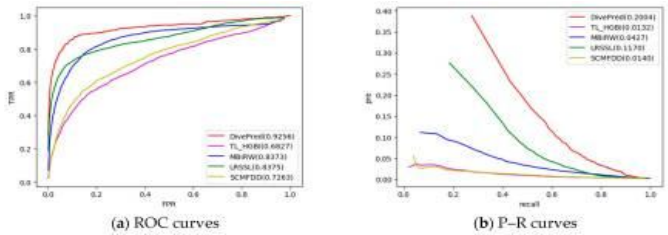
$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{TN + FP} \quad [2]$$

Where TP is the number of positive cases that were correctly identified, and TN indicates the number of negative examples that were correctly identified. FN and FP are the numbers of positive and negative examples that were misidentified, respectively. After calculating TPRs and FPRs for different θ values, the receiver operating characteristic curve (ROC) was be plotted. The area under the curve (AUC) was used as a measure to predict the performance of potentially associated disease with drug ri. The overall performance of the prediction method was the average of the AUC values of all drugs.

Due to the imbalance of the number of positive and negative samples in the sample data, the precision–recovery rate (P–R curve) can provide additional information; precision and recall were defined as follows,

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{TP + FN} \quad [2]$$

The precision ratio refers to the proportion of correctly identified positive samples in the search samples, and the recall rate is the same as the TPR. The area under the P–R curve (AUPR) was also used to measure the performance for predicting potential drug–disease associations. [2]



(a) ROC curves (b) P–R curves

Fig. 3 Two types of curves for evaluating the predicting performance of DivePred and other methods. (a) receiver operating characteristic (ROC) curves; (b) precision–recall (P–R) curves. [2]

©IJRASET: All Rights are Reserved

Biologists normally pick the highest-level possibility for additional experimentation. It was our objective to build the quantity of positive examples in the highest-level area. To make another assessment record, we determined the review pace of the highest-level examples, which is the extent of positive examples accurately recognized in the top k of the rundown among the complete of positive examples.

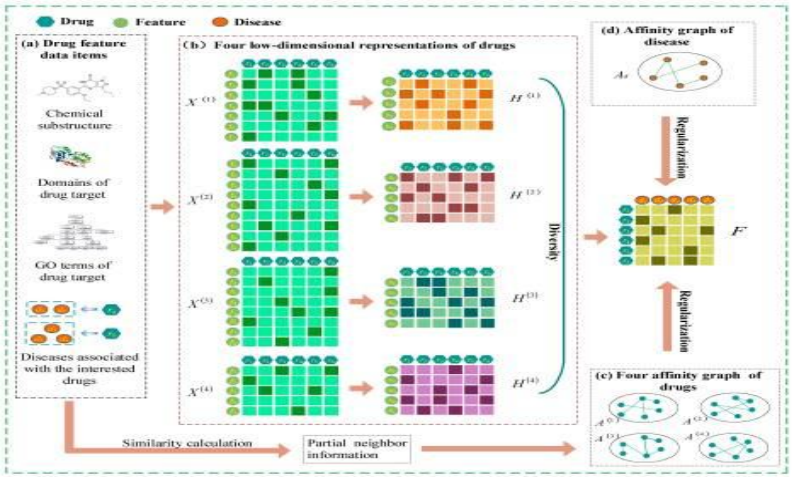


Fig 4. Representation of data from drugs and diseases from multiple sources and representation of drug–disease predictive association matrix F. (a) Drug feature data sets from multiple sources; (b) four low-dimensional representation of drugs; (c) four affinity maps of the drugs were obtained by similarity calculation; (d) extract the similarity of the diseases and obtain the affinity map of the disease.[2]

5) *Benefits:* DivePred outperforms other methods in AUCs and AUPRs. For biologists, DivePred is very useful because more real drug–disease associations were included in DivePred's top-ranking candidate list. Case studies on five drugs demonstrated that DivePred could detect potentially new indications for drugs.
6) *Scope:* DivePred can fill in as a prioritization apparatus to screen the likely contender for resulting disclosure of genuine medication illness relationship through natural approval.

III. PROPOSED WORK

To begin with, the data set serves a base that provides insights about the current system which talks regarding Medicine's sales details like Medicine name, Sales quantity.

In the research project, the proposed system works on a model which takes data from the dataset [4][5] of the current system, analyses the data with the help of some additional variables such as Medicine Name, Users Feedback or Rating and Condition for which the medicine is used. From the analysis, the proposed system derives the sales quantity of the medicine and predicts diseases associated with the medicine along with the effectiveness and ratings based on respondent's feedback.



Fig 5. Working of the system

A. *Technial Information*

1) *Technologies Used*

a) Streamlit Framwork
b) Python as an language

- Numpy package
- Plotly package
- Nltk Package
- Pandas Package

©IJRASET: All Rights are Reserved

2) Minimum System Configuration

a) Hardware

- Quad core processor
- 4GB Ram
- 100 GB Storage

b) Software Configruation

- Windows / MAC / Linux OS
- Python Environment
- Streamlit

B. Data Set

The data set used for analysis in the model consists of the following variables

1) Medicine Sales Details – Medicine name, sales quantity Along with the above dataset few additional variables play an important role in analysis which are as follows

2) Medicine and Disease Details – Medicine name, Associate diseases or Condition, Reviews, Usefulness

C. Data Summary

In this section, the information of the dataset used in the proposed system has been elaborated

We have 2 datasets (A and B)

Dataset "A" – Medicine Sales data consisting information about the sales quantity and medicines names.

The data set has total 4,64,965 entries in which the total quantity of sales is 8,62,82,849 of 4672 Medicines

Dataset "B" – The dataset consists information about Medicine, Disease / condition associated with the medicine, effectiveness and reviews based on users feedback.

Furthermore, to be precise this dataset has in total 53,766 entries with 2,637 unique data of Drugs / Medicine with 708 Disease / Conditions, 48,280 reviews and 53,766 ratings taken from respondents.

D. Word Frequency

This section talks about the words used maximum by respondents in the survey

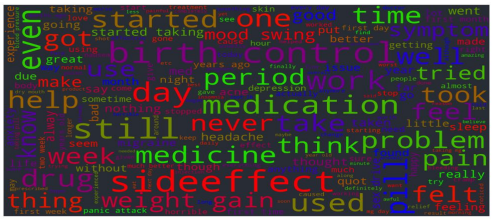


Fig 6. Word Cloud

As per the figure, as stated above it can be seen that the maximum times words used by respondents in the survey are Side-effect, medication, Birth control, problem, pain, drug, etc.

E. Sentiment Analysis

This section talks about the sentiments of the respondents based on the feedback given by them on the survey. The sentiments of the users are derived from the feedback given by them in the survey using machine learning technique, through the analysis we get the results of their sentiments in 3 categories: Positive, Negative, Neutral as mentioned below:



Fig 7. Sentiment count

©IJRASET: All Rights are Reserved

F. Medicine Analysis

This section talks about analysis performed by the system based on the list of medicine available from the data set and displays information about diseases associated with any particular medicine selected by the user along with the total count that represents maximum usage of the medicine from the respondent's feedback. The analysis also displays information about the effectiveness of the medicine on the diseases associated with it and also gives rating (0-10) based on the feedback from the respondents.

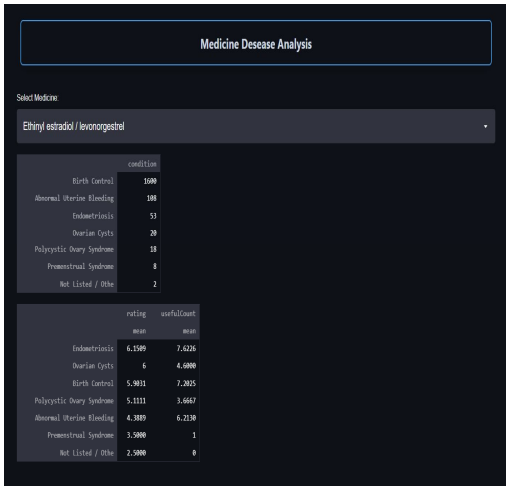


Fig 8. List of dieases associated with the selected medicine

G. Disease Analysis

This section talks about analysis performed by the system based on the list of diseases available from the data set and displays information about medicine associated with any particular disease selected by the user along with the total count that represents maximum occurrence of diseases as per respondent's feedback. The analysis also displays information about the effectiveness of the medicine on the diseases associated with it and also gives rating (0-10) based on the feedback from the respondents

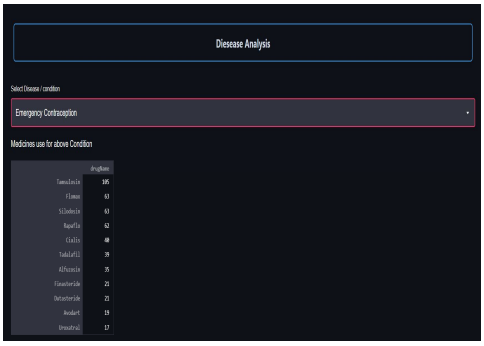


Fig 9. List of Medicines associated with a Disease

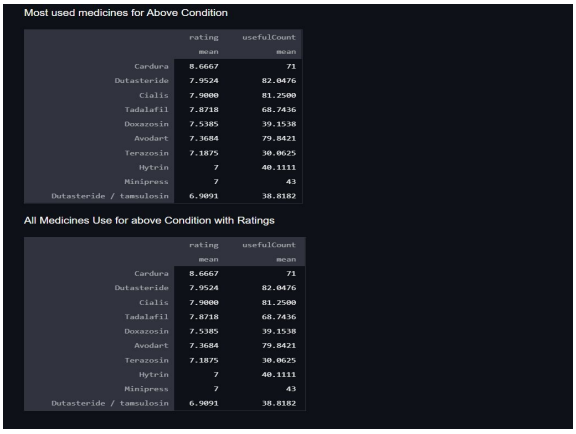


Fig 10. List of medicine and Ratings given by the particiapants

©IJRASET: All Rights are Reserved

H. Sales Data Analysis

This section talks about the sales data obtained from dataset A, where the list of unique medicines is identified and displayed. Post that top 10 medicines are identified according to the sum of total quantities of medicines sold. the system compares list of top 10 medicines data with dataset B consisting of medicine name and diseases associated with it. Now the system displays two lists where we get information about medicine used for particular diseases and other associated diseases form dataset B as per the top 10 sales records from system.



Fig 11. List of Medicine Present In Dataset

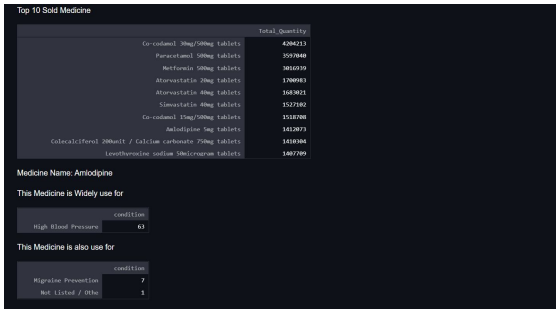


Fig 12. Top 10 Medicine which sold and Medicine information Like Widely used for Dieases and another associate dieases.

IV. CONCLUSION

In the current medical system, the sales data is only used to analyses the sales of Medicines which does not inform about the effectiveness of any medicine along with the list of diseases it can be helpful for curing.

For the research project, two literature reviews were conducted to understand the relation between medicines and diseases. Furthermore, the first literature review has informed us how a molecule plays an important role in medicine to cure any disease in a human body. Whereas, the second literature review stated that any disease could be treated with the help of one or more medicines. Similarly, we can learn that any medicine could be used to treat more than one disease.

In the research project, the proposed system works on a model which takes information from the dataset of the current system, investigations the information with the assistance of

some extra factors, for example, Medicine Name, Users Feedback or Rating and Condition for which the medication is utilized. From the examination, the proposed system infers the sales quantity of the medication and predicts sicknesses related with the medication alongside the viability and appraisals dependent on respondent's feedback.

V. LIMITATIONS

In the research project, it can be said that there are three limitation of the system which are crucial for analysis. These limitations are - Dataset, Diseases association, Participation.

1) *Dataset* – For apt analysis, data set plays a major role in the system. Unorganized information forms the dataset can lead to improper results, hence every information from dataset such as Medicine Name and Sales Quantity should be present with correct association of medicine with the disease.
2) *Diseases Association* – In order for the system to analyse and predict correct association of medicine with disease, information about list of medicine and diseases should be correctly present.
3) *Participation* – According to the secondary research undertaken, the research project observes feedback of various respondents sharing insights about various medicine that helped them in the treatment of their disease. For more optimised analysis, the information about respondents' age and gender would be more beneficial.

©IJRASET: All Rights are Reserved

VI. SCOPE

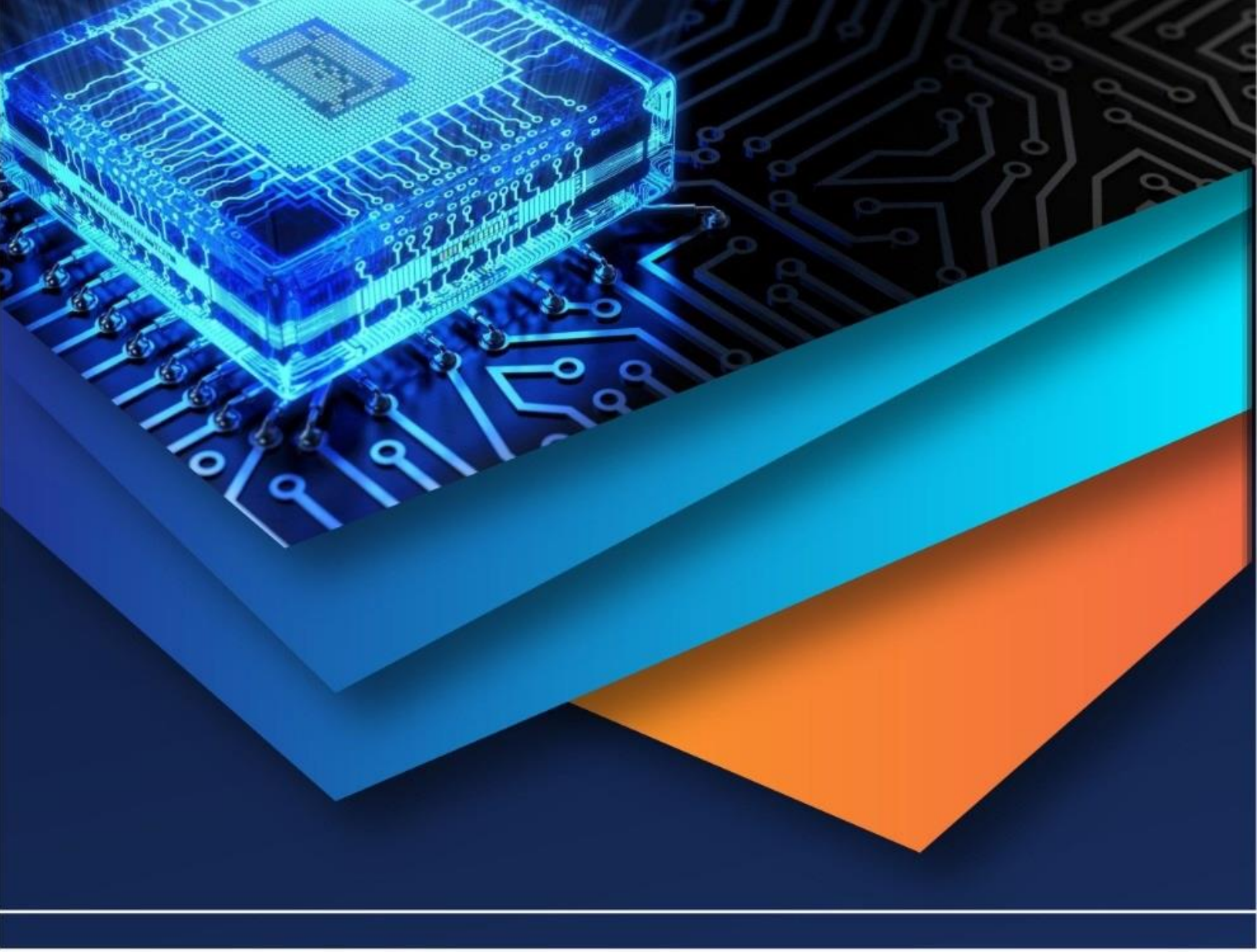
The research project describes a model / system that displays a list of diseases on selecting medicine name which is useful for treating the disease, also its displays the list of medicine that can be used for treating a particular disease by selecting the diseases name. Here, the advantage is that user can check with the system and get an idea about the medicine and disease associated with it and further he/she can check with the medical professionals for further guidance.

The scope of the project can be enhanced on using the system where the sales data can be obtained from local pharmacist and on getting this data, people can check which medicines are associated with one or more diseases and vice-versa. Using this information from system, health departments of various small areas can check which medicine is sold more for treatment of diseases, and observing a hike in sales of any medicine the health department can track which area is suffering from any disease and likewise they can take precautionary measures for the same.

REFERENCES

[1] Mathilde Pivette, Judith E Mueller, Pascal Crepey, Anver Bar-Hen "Drugs sales analysis for outbreak detection of infectious Diseases" BMC Infect Dis 18 Nov 2014 edition 14: 604
[2] Ping Xuan, Yingying Song, Tiangang Zhang, Lan Jia "Prediction of Potential Drug-Disease Associations Through Deep Integration of Diversity and Projections of Various Drug Features", Int J Mol Sci, 20 Sep 2019 (4102).
[3] Typical Supply Chain Structure fig.1. http://dolcera.com/wiki/index.php/Indian_Pharma_Industry_-_Distribution_&_Sales_Force_Structure. .
[4] Dataset A: https://data.world/fivethirtyeight/study-drugs
[5] Dataset B: https://ckan.publishing.service.gov.uk/dataset/gp-prescribing-data

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)