Conversational AI: Accelerated Data Science REPORT

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

(Deemed to be University), Patiala – 147004

Submitted By:

Mohammed Ayesha Sami (102117110)

Pragya Gupta (102103407)

Submitted To:

Mr. Arun Pundir

ABSTRACT

Large data blocks must be analysed and explored using the data mining technique to identify relevant trends and patterns. Medical databases are one of the many fields in which data mining techniques can be used. Thousands of people worldwide are dealing with health and issues with medical diagnosis. Hospital Information Systems (HIS) produce vast amounts of data, but extracting meaningful information from diagnosis case data can be challenging. By simply their condition, patients can quickly get information about the the medication that can help treat it, thanks to the techniques used in this project.

In this project, we provide drug recommendations to users based on reviews and useful count.

The reviews are analysed using the Vader tool and NLP-based sentiment analysis. Finally, the drugs are recommended using weighted average approaches.

INTRODUCTION

a) Problem description:

When confronted with any medical condition, one of the most common concerns among patients is which physician to trust. According to the Times of India, 5.2 million people in India die each year due to medication errors. More than 42% of medication errors are caused by doctors because they write prescriptions based on their experience, which is quite limited. Hence, finding appropriate physicians to diagnose and treat medical conditions is one of the most important decisions a patient must make. Advancements in data mining and Recommender Technologies allow us to explore possibilities for potential knowledge from diagnosis history records and reviews and ratings on drugs to help doctors prescribe the correct medication and decrease medication errors effectively.

b) Problem challenges:

The challenges inherent in the current healthcare landscape are multifaceted and demand innovative solutions. Firstly, the sheer volume of medical data within Hospital Information Systems (HIS) poses a significant hurdle. Extracting actionable insights from these vast datasets requires sophisticated data mining techniques capable of identifying relevant trends and patterns amidst the noise.

Moreover, the complexity of medical diagnoses compounds the difficulty. With thousands of people worldwide grappling with health issues and medical diagnosis, the need for accurate and timely information is paramount. Traditional methods of diagnosis and treatment often rely on subjective factors, such as a physician's experience or intuition, which can lead to errors and suboptimal outcomes.

c) Novelty in work:

This project introduces an innovative approach to drug recommendation by integrating sentiment analysis of drug reviews with weighted average rating calculation. Leveraging Natural Language Processing (NLP) techniques, particularly the VADER tool, our method accurately assesses the sentiment polarity and intensity of drug reviews, ensuring that only medications with positive user feedback are considered for recommendation. By dynamically scraping data from drugs.com and synthesizing multiple data streams, including numerical ratings and useful counts, our recommendation pipeline offers a comprehensive solution for personalized drug recommendations tailored to individual patient conditions. This patient-centric approach prioritizes treatment efficacy and patient satisfaction, aiming to revolutionize medication selection and treatment planning in healthcare.

LITERATURE SURVEY

[3]	Luis Fernando Granda; Priscila Valdiviezo-Diaz; Ruth Reátegui & Luis Barba-Guaman (2022)	The study used patient data from UC Irvine's Machine Learning Repository on diabetes, applying data mining for clustering and dimensionality reduction. Drug predictions were made via collaborative filtering, matching similar patient profiles for personalized recommendations. System performance was evaluated for group quality and prediction accuracy, aiming to enhance drug recommendation systems for diabetes patients.	Collaborative Filtering and Clustering Approach
[1]	Kamaraj, K.Gomathi & Priyaa, D.Shanmuga. (2016)	In this paper, they have demonstrated the results of Decision Tree and Naive Bayes models in predicting three diseases, Heart Disease, Diabetes, and Breast Cancer.	Decision trees, Naïve Bayes model
[2]	M.A.Nishara Banu, B Gomathy. (2013)	In this paper, they have predicted heart diseases. By applying the KMean on the medical dataset, they have clustered the relevant data, upon which MAFIA(Maximal Frequent Itemset Algorithm) is applied to generate rules and identification of frequent patterns, which is fed to the C4.5 (Decision Tree) model to classify patterns.	K-means. MAFIA
[4]	Abhishek, Amit Kumar Bindal, Dharminder Yadav	In this paper, they use a medicine recommendation system using the R programming language and the recommended lab package. It utilizes actual user ratings to predict top medicine choices and analyzes user review sentiment toward medicines.	Collaborative filtering, Contentbased filtering

METHODOLOGY

The main goal of our project is to recommend a drug to a patient based on the condition she/he has. In accordance with our objective to implement a drug recommender system there is one main category which is to be addressed i.e. a drug recommender model and below is the design pipeline and dataflow of our implementation.

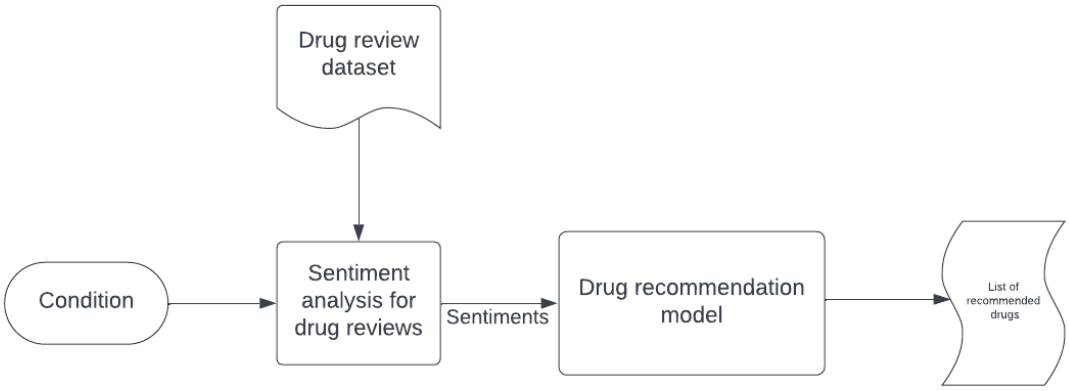


Figure 1: Design implementation pipeline and dataflow of our recommender system. The sentiment analysis on the drug reviews is mapped to condition and finally recommended drugs based on the sentiments of the reviews and rank.

Sentiment Analysis of drug reviews

Our task is to be able to recommend the best drug for the patient. For this we have used to NLP approach to analyse the sentiments using a VADER tool. It is a simple rule-based model for general sentiment analysis. The VADER stands for Valence Aware Dictionary for Sentiment Reasoner. It is a lexicon and rule-based sentiment analysis tool that is specifically tuned to sentiments expressed in social media. This model does not require any training data as it already uses a combination of qualitative and quantitative methods to produce and then empirically validate a gold-standard sentiment lexicon to evaluate the sentiment of the sentence. This tool will take a string input which will be a drug review from our dataset. This tool not only tells us about the polarity (positive or negative) of the review, but also how much positive or negative the review is.

There are many advantages of using this tool. They include:
1. It works exceptionally on the social media style texts which may include emoticons and punctuations.
2. It requires no training data.
3. Can be used with real time streaming data.
4. It does not suffer from a speed-performance trade-off.
The output generated has 4 components i.e. Positive, Negative, Neutral and Compound.
Positive, Neutral and Negative specify how many parts of the sentence have the tone as positive, neutral, and negative respectively. It is specified in the form of decimal numbers and the sum of these three components will always be equal to 1. The compound component specifies the overall sentiment of the sentence.

Below thresholds help us to interpret results of compound:
1. Positive sentiment: compound score >= 0.05
2. Neutral sentiment: -0.05 < compound score < 0.05
3. Negative sentiment: compound score <= -0.05

You can see the output of a sample review in the below figure. So here in the sentence, there is a bit of positive part in the sentence, but the overall sentiment can be computed as negative which is evident by the compound value given by VADER.

```
[57] #Using a sample review to see the output of VADER tool.

     sentence = "This medicine is really helpful with voices but my paranoia and depression has got worse"
     analyser.polarity_scores(sentence)

     {'compound': -0.8923, 'neg': 0.473, 'neu': 0.445, 'pos': 0.083}
```

Drug Recommendation

After we have categorized the reviews into positive, negative, and neutral, we need to recommend a drug for the condition. For recommendation of the drug, we have scraped the data from drugs.com and created a review dataset which has the condition along with its multiple available drugs, their reviews, ratings and useful count.

There are 6 features in this dataset out of which 3 are the ones that will help us to recommend the best drug. These 3 features and their importance is explained below:

● Feature 1. Review:
The review column is one of the most important columns for drug recommendation as it is direct feedback from the users after using that drug. Hence, by doing Sentiment Analysis, we have taken into consideration only the reviews with a positive sentiment.

● Feature 2 & 3. Rating & Useful Count:
Every drug has a rating associated with it from 1-10 and a useful count which tells us how many users have given that rating. This information can be used to find a weighted average rating.

Weighted Average Approach:

Weighted average or mean is like an ordinary arithmetic mean, except that instead of each of the data points contributing equally to the final average, some data points contribute more than others.

Formula for weighted average:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

which expands to:

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}.$$

where, data elements with a high weight contribute more to the weighted mean than do elements with a low weight.
In our case, x is the rating and w is the useful count. In this way, we get a weighted average rating which gives more importance to the useful count i.e. the number of people who have, given that rating.

This is the function created for calculating weighted average:

```
def wavg(group, avg_name, weight_name):
    d = group[avg_name]
    w = group[weight_name]
    try:
        return (d * w).sum() / w.sum()
    except ZeroDivisionError:
        return d.mean()
```

```
data_wavg = data.groupby(["Drug"], as_index=False).apply(wavg, "Rating", "UsefulCount")
```

After calculating, this weighted average column is merged with the dataset:

	drugName	Rating_Wavg	condition	review	Review_Sentiment	rating	usefulCount
0	Voltaren	8.854352	zen Shoulde	"Great help"	Positive	8	33
4	Diclofenac	8.617426	zen Shoulde	"Great help"	Positive	8	33
5	Relafen	8.442509	zen Shoulde	"I'm probably the only one I know taking ...	Positive	7	50
6	Ibuprofen	8.428997	zen Shoulde	"I've found that taking ibuprofen (200 mg...	Positive	8	0
8	Naproxen	7.947768	zen Shoulde	"Very little relief. I finished PT and after ...	Positive	2	6

Table 1: After weighted average calculation.

- This merged dataset now has the condition and drug name along with the review and its sentiment and the average rating with useful count.
- For recommending drugs to the user, we have only considered drugs which have a positive sentiment in their reviews. Hence, filtering out unwanted drugs.
- Also, we have taken a total of all the useful counts present for a particular drug to sort the drug that must be recommended with the highest useful count first and then the highest average rating.
- For this, we have used the groupby.sum() method as shown below:

```
# taking predicted disease as input and recommending drug based on highest weighted average of ratings
groupedByCount = merged_wavg.groupby(['Disease', 'Drug', 'Rating_Wavg'])['UsefulCount'].sum().reset_index
()
```

This approach allows us to use all the information present in the dataset according to its importance and hence as an output we get the best possible drug recommendation.

RESULT AND ANALYSIS

DATASET DESCRIPTION:

Dataset Gathering

● This dataset is used in order to take the condition as input and recommend appropriate drugs based on reviews and ratings (Sentiment Analysis).
● The dataset is scraped from Drugs.com (https://www.drugs.com/) for Drug Review which provides patient reviews and useful count which describes how many users find it helpful along with related conditions and a 10-star patient rating reflecting overall patient satisfaction.
● It contains 6 columns and 161297 rows:

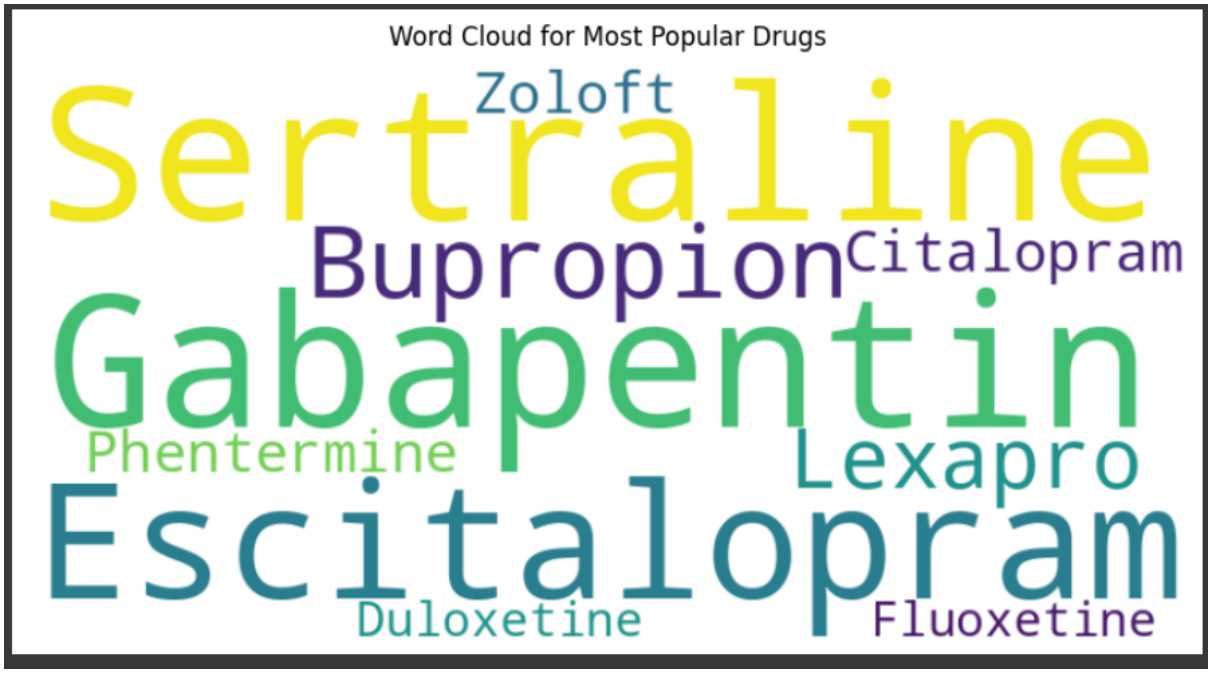
1. ID
2. Drug name
3. Condition
4. Review
5. Rating
6. Date
7. Useful count

```
[ ] df.head()
```

	drugName	condition	review	rating	usefulCount
0	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	27
1	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	192
2	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	17
3	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	10
4	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	37

●The dataset obtained was clean and did not require a lot of pre-processing. Still, a few rows with null values were dropped and columns were renamed.
● The dataset contains a lot of information and visualizing was an interesting task.

● One such visualization is shown below which has names of a few most popular drugs:



The VADER analysis approach:
To get the sentiment analysis of each drug review, we have used vader analysis. The output given by vader consists of 4 components namely, positive, negative, neutral, and compound. After doing the vader analysis of every review, below we have shown results for every review in the dataset.

```
[ ] #Visualizing data from vader analysis.
    sentiments_data
```

	Review	Positive	Negative	Neutral	Compound
0	"It has no side effect, I take it in combinati...	0.000	0.121	0.879	-0.2960
1	"My son is halfway through his fourth week of ...	0.108	0.018	0.874	0.9174
2	"I used to take another oral contraceptive, wh...	0.080	0.059	0.861	0.6160
3	"This is my first time using any form of birth...	0.089	0.026	0.885	0.7184
4	"Suboxone has completely turned my life around...	0.168	0.061	0.771	0.9403
...	...	...	...	...	...
161292	"I wrote my first report in Mid-October of 201...	0.143	0.037	0.820	0.9366
161293	"I was given this in IV before surgey. I immed...	0.086	0.139	0.775	-0.4767
161294	"Limited improvement after 4 months, developed...	0.135	0.462	0.404	-0.7430
161295	"I've been on thyroid medication 49 years...	0.113	0.066	0.821	0.8503
161296	"I've had chronic constipation all my adu...	0.179	0.026	0.794	0.9043

161297 rows × 5 columns

Figure 2: Vader Sentiment analysis for all reviews.

To classify the overall sentiment of every review, we used the threshold specified earlier. After assigning each review its overall sentiment, we can see distribution of positive, negative, and neutral reviews in the dataset in the below diagram.

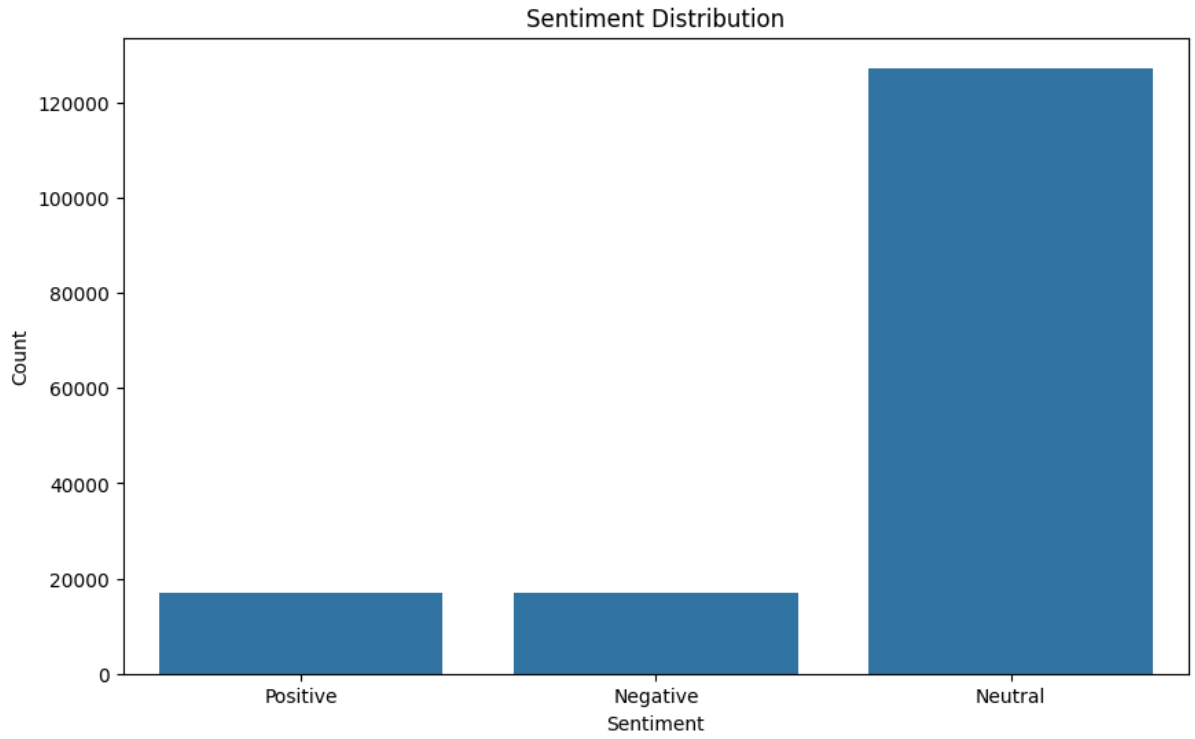


Figure 3: Graph showing sentiment of reviews in the dataset.

Drug Recommendation

Few things considered while recommending the drug:

● There are multiple drugs for a single condition. Hence, by using sentiment analysis we were able to filter out the negative and neutral reviews leaving us with only positive ones.

Weighted Average Approach

The results obtained from this approach gave us the correct and highly rated drug for the condition. Based on the given condition, we are recommending top 3 drugs from our dataset as an output for Insomnia as input. It can be seen below:

```
[83] recommended_drug = pd.DataFrame(groupedByCondition.get_group('Insomnia').nlargest(3, ['Rating_Wavg', 'usefulCount']))
     recommended_drug
```

	condition	drugName	Rating_Wavg	usefulCount
3522	Insomnia	Secobarbital	10.000000	35
3514	Insomnia	Pentobarbital	10.000000	27
3504	Insomnia	Intermezzo	9.770492	122

Output of weighted average approach for drug recommendation

- Verification of results with real world

After doing a google search and selecting the trusted resources for providing medical health information like MedlinePlus (https://medlineplus.gov), Drugbank (https://go.drugbank.com) and U.S. FOOD & DRUG (https://www.fda.gov/) on the drugs recommended by the model namely Secobarbital, Pentobarbital and Intermezzo, we have found the below results to see if they match the condition.

Verifying results for the condition Insomnia from the above-mentioned websites:

S.NO	RECOMMENDED DRUG	VERIFYING RESULTS
1	Secobarbital	Secobarbital is used on a short-term basis to treat insomnia
2	Pentobarbital	Pentobarbital is a barbiturate drug used to induce sleep, cause sedation
3	Intermezzo	For the short-term treatment of insomnia

Verified screenshots are attached below:

Why is this medication prescribed?

Secobarbital is used on a short-term basis to treat insomnia (difficulty falling asleep or staying asleep). It is also used to relieve anxiety before surgery. Secobarbital is in a class of medications called barbiturates. It works by slowing activity in the brain.

Pentobarbital

Summary Pentobarbital is a barbiturate drug used to induce sleep, cause sedation, and control certain types of seizures.

Indication	For the short-term treatment of insomnia.

----------INDICATIONS AND USAGE----------

Intermezzo is a $GABA_A$ agonist indicated for use as needed for the treatment of insomnia when a middle-of-the-night awakening is followed by difficulty returning to sleep (1)

Hence, we can say that our drug recommendation system works as expected and gives the desired results!

CONCLUSION AND FUTURE SCOPE

Drug recommendation systems represent a prevalent technology in contemporary online services. As the demand for these services escalates, a pressing need for automation arises. Below, we outline the key findings of our project:

- Successfully devised a drug recommendation system capable suggesting medications, based on user-input conditions.
- Implemented one distinct model for the project: a sentiment analysis model.
- Achieved commendable accuracies with the model used, thereby enhancing the overall reliability of the drug recommendation system.

One essential future scope can be improving the accuracy of the prediction and recommender model using deep neural networks by using larger data.

REFERENCES

[1] Kamaraj, K.Gomathi & Priyaa, D.Shanmuga. (2016). Multi Disease Prediction using Data Mining Techniques. International Journal of System and Software Engineering.

[2] M.A.Nishara Banu, B Gomathy. (2013). Disease Predicting System Using Data Mining Techniques. International Journal of Technical Research and Applications.

[3] Granda Morales LF, Valdiviezo-Diaz P, Reátegui R, Barba-Guaman L. Drug Recommendation System for Diabetes Using a Collaborative Filtering and Clustering Approach: Development and Performance Evaluation. J Med Internet Res.

[4] Yin Zhang, Daqiang Zhang, Mohammad Mehedi Hassan, Atif Alamri & Limei Peng (2014). CADRE: Cloud-Assisted Drug REcommendation Service for Online Pharmacies.

[5] Youjun Bao ; Xiaohong Jiang. (2016). An intelligent medicine recommender system framework.

[6] Druglib.com - Drug Information, Research, Clinical Trials, News. http://www.druglib.com/

[7] Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014

[8] H Wang Q Gu J Wei Z Cao Q Liu (2015). Mining drug–disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies.

[9] R.J. Mooney, P.N. Bennett, and L. Roy, In: Proc. Recommender Systems Papers from 1998 Work-shop,
Technical Report, 49-54 (1998).
[10] C. Silpa, B. Sravani, D. Vinay, C. Mounika and K. Poorvitha, "Drug Recommendation System in Medical Emergencies using Machine Learning

[11] Doulaverakis, C., Nikolaidis, G., Kleontas, A., and Kompatsiaris, I. 2012. GalenOWL: Ontology-based drug recommendations discovery. Journal of Biomedical Semantics

[12] Zhang, Y., Zhang, D., Hassan, M. M., Alamri, A., and Peng, L. 2015. CADRE. Cloud-Assisted Drug recommendation Service for Online Pharmacies.
[13] S. Garg, "Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning," 2021

[14] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. Practical Journal of Clinical Medicine, 4.

[15] Bhimavarapu, U.; Chintalapudi, N.; Battineni, G. A Fair and Safe Usage Drug Recommendation System in Medical Emergencies by a Stacked ANN. Algorithms 2022