# AI based Medical Diagnosis with Medicine Recommendation

Janvi Patel
tuo93411@temple.edu

## *Abstract –*

The proposed idea focuses on the concept of decision support system to support doctors and act as a validation tool in the process of diagnosis and predicting disease along with prescribing the proper medications according to the disease detected. However, diagnosing disease can be a tricky task as there are many common symptoms related to different diseases and accurately predicting the disease is very important here. This project consists of two phases in 1st phase a supervised learning approach is used for the prediction of the diseases. Various ML models have implemented, and the model which will give the best accuracy will be used for further processing. In the 2nd phase a recommendation system is implemented which will take the predicted disease from the phase 1 and recommended the appropriate medicines based on their ratings. Sentiment analysis is used for identifying which ratings are positive, negative, and neutral.

## *Introduction –*

Accurate diagnosis is a fundamental aspect of global healthcare systems. According to the statistics in [1] approximately 5% of outpatients in US receives an incorrect diagnosis, with errors being particularly common for serious medical conditions carrying the risk of serious patient harm. The conventional method used for diagnosing disease is somewhat error prone. Using the AI technique to diagnosis the disease reduces the chances of errors and increases the accuracy of identifying the disease based on the symptoms and using the knowledge. There are many Machine Learning Techniques like Supervised, Unsupervised, Deep Learning, etc. which are used to classify huge data very efficiently. So, I have used K-Nearest Neighbor and Support Vector Classifier with GridSeacrchCV algorithms for accurate classification and prediction of disease.

The disease classification and medicine recommendation are a challenging task because there are many symptoms which are common for different disease. Drug rating and review are key attributes in medicine recommendation systems. Here I first used the supervised machine learning algorithm to predict the disease based on symptoms and then build a recommendation system for medicine and then finally combined both the methods to predict the disease and then based on that predicted disease recommend the highest rated medicine for that disease. Following sections are as follows: related work, followed by the proposed methodology, then comes the results and then discussion which concludes the work done.

## *Related Work –*

Strong AI techniques can unlock clinically relevant information hidden in the massive amount of data guided by relevant clinical questions, which in turn can help clinical decision-making [3]. In [2] the formulated approach used for recommending medicines was based on the ratings count. The disease prediction approach in [2] uses the disease occurrence i.e., count of the disease as one of the parameters for making the predictions. The highest accuracy obtained by [2] was 92% which uses the decision tree model.

## *Methodology –*

Through my project, this research propose an AI system which predicts the disease based on the symptoms followed by recommending the appropriate best medicine based on their ratings [2]. Fig. 1 shows the basic architecture of the system. This section has 2 sub-sections: Disease prediction and Medicine Recommendations.

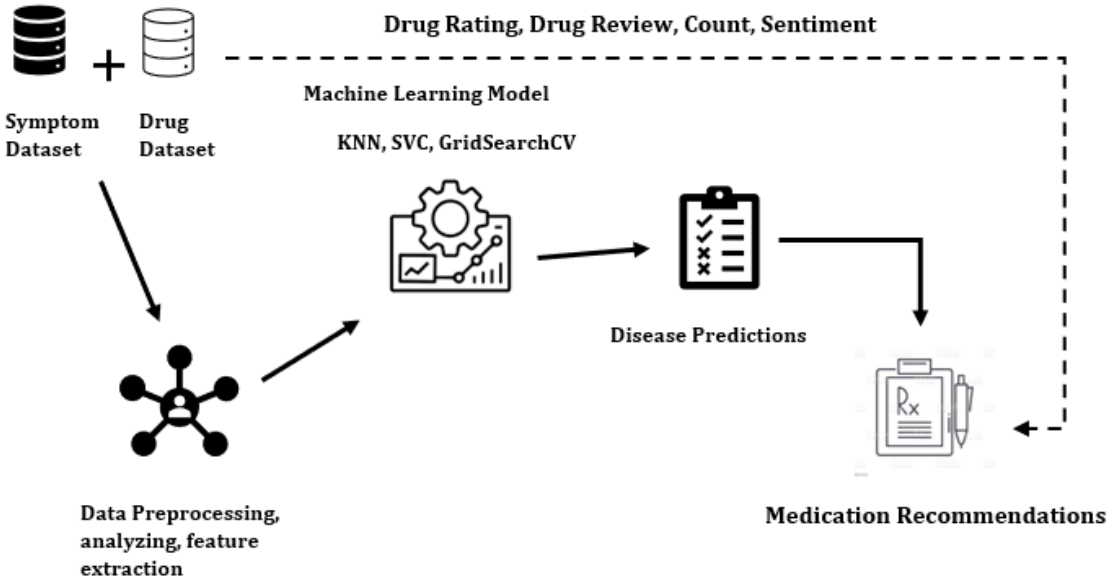


*Fig. 1: System Architecture*

1. ***Disease Prediction –*** For the prediction of diseases, the multiple datasets were collected from the Kaggle platform. A final dataset was created for the final analysis of the system. The first dataset which was collected consists of 1 disease column and 17 symptoms columns. Fig. 2 shows the disease-symptom dataset. There were 41 unique disease presents in the dataset. And based on the preprocessing and the correlation between the features only 8 symptoms were chosen. Fig. 3 shows the heatmap generated for the correlation matrix for finding the relationship between the symptoms.

| | Disease | Symptom_1 | Symptom_2 | Symptom_3 | Symptom_4 | Symptom_5 | Symptom_6 | Symptom_7 |
|---|---|---|---|---|---|---|---|---|
| 0 | Fungal infection | itching | skin_rash | nodal_skin_eruptions | dischromic _patches | 0 | 0 | 0 |
| 1 | Fungal infection | skin_rash | nodal_skin_eruptions | dischromic _patches | 0 | 0 | 0 | 0 |
| 2 | Fungal infection | itching | nodal_skin_eruptions | dischromic _patches | 0 | 0 | 0 | 0 |
| 3 | Fungal infection | itching | skin_rash | dischromic _patches | 0 | 0 | 0 | 0 |
| 4 | Fungal infection | itching | skin_rash | nodal_skin_eruptions | 0 | 0 | 0 | 0 |

*Fig. 2: Disease-Symptom dataset*

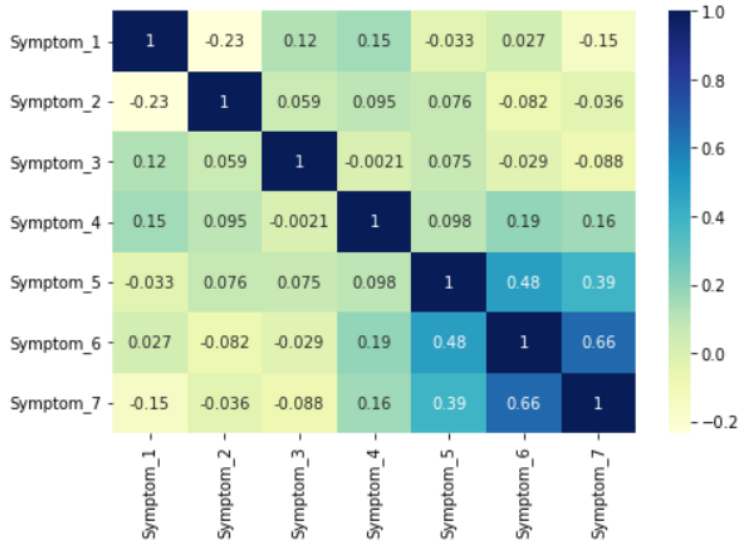


*Fig. 3: heatmap generated for correlation matrix*

The dataset shown in fig.2 was then merged with the severity dataset. Where the symptoms of the diseases are organized based on their severity weight. High severe symptoms were given high value and low severe symptoms for a disease were given lower value. Fig. 4 shows the final dataset for disease prediction.

| | Disease | Symptom_1 | Symptom_2 | Symptom_3 | Symptom_4 | Symptom_5 | Symptom_6 | Symptom_7 |
|---|---|---|---|---|---|---|---|---|
| 0 | Fungal infection | 1 | 3 | 4 | 0 | 0 | 0 | 0 |
| 1 | Fungal infection | 3 | 4 | 0 | 0 | 0 | 0 | 0 |
| 2 | Fungal infection | 1 | 4 | 0 | 0 | 0 | 0 | 0 |
| 3 | Fungal infection | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 4 | Fungal infection | 1 | 3 | 4 | 0 | 0 | 0 | 0 |

*Fig. 4: Final dataset for disease prediction*

Fig. 5 shows the top 10 diseases in the dataset. For disease predictions 3 machine learning models were trained. The first was the KNN model. Since k nearest neighbor works on the principle of proximity i.e., how similar the data points are this model helps to effectively predict the disease as the symptoms are closely related to each other of a disease. For this model distance measure between symptoms is calculated and only those symptoms which have smaller distances is grouped for predicting the associate disease. For finding the best value of K an error rate for K values was plotted. Fig. 6 shows the Error rate K value plot. The second model which was trained was the Support Vector Classifier model. SVM tries to classify cases by finding a separating boundary called hyperplane. SVC has recently been used to improve methods for detecting diseases in clinical settings. SVC is one of the powerful supervised algorithms which works best on small datasets but on complex ones.

The SVC model was then hypertuned using the GridSearchCV to find out if it gives better results compared to other models.

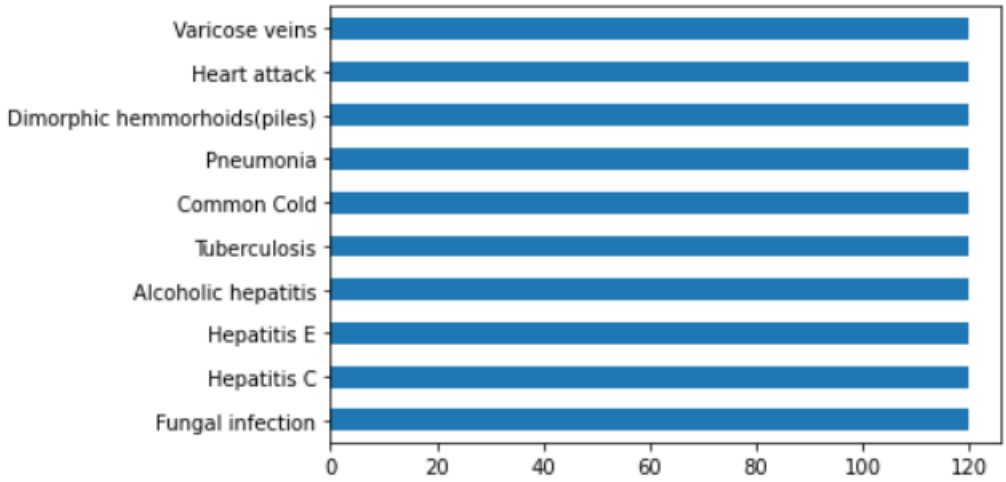


*Fig. 5: top 10 diseases*

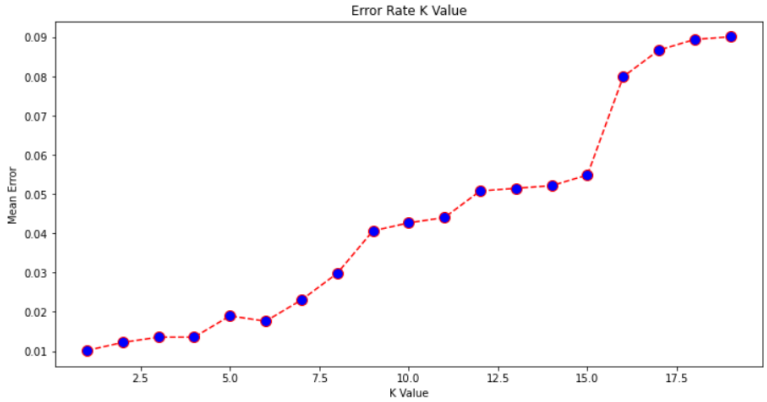


*Fig. 6: Error rate K value*

2. ***Medicine Recommendations*** **–** For the recommendation's sentiment analysis was done on the reviews of the medicines. Only the positive sentiments were used for final recommendations. The dataset used for recommendations was the drug-disease dataset collected from Kaggle. The drug-disease dataset was combined with disease-symptom dataset for getting the similar diseases from both the datasets. Fig. 7 shows the combined dataset. SentimentIntensityAnalyzer was used on the review column for getting the polarity scores i.e. positive, negative, neutral and compound for the reviews. Out of which only positive and negative sentiments were added to the dataset for the recommendation step. After analyzing the sentiments groups of diseases and medicines were formed based on their ratings and the count. Fig. 8 shows the groups of diseases and medicines. Which was then added to the dataset as weighted average ratings. Fig. 9 shows the final dataset used for the recommendation stage.

| Drug | Disease | Review | Rating | UsefulCount | Symptoms |
|---|---|---|---|---|---|
| Sulfamethoxazole / trimethoprim | urinary tract infection | "I have bad side effects from all antibiotics,... | 1 | 17 | [' burning_micturition', ' bladder_discomfort'... |
| Levofloxacin | urinary tract infection | "This last Monday I was detected by my doctor ... | 3 | 24 | [' burning_micturition', ' bladder_discomfort'... |
| Sulfamethoxazole / trimethoprim | urinary tract infection | "I am currently suffering from recurring cysti... | 1 | 7 | [' burning_micturition', ' bladder_discomfort'... |
| Nitrofurantoin | urinary tract infection | "I have been taking 4 a day for last 6 days, 2... | 5 | 10 | [' burning_micturition', ' bladder_discomfort'... |
| Macrobid | urinary tract infection | "I wish I would have read these reviews before... | 1 | 28 | [' burning_micturition', ' bladder_discomfort'... |

*Fig.7: Combined medicine-disease-symptom dataset*

```
Disease                Drug
acne                   Absorica                          6.692308
                       Acanya                            7.000000
                       Accutane                          8.798081
                       Acnex                            10.000000
                       Aczone                            8.584971
                                                           ...
urinary tract infection  Sulfamethoxazole / trimethoprim  5.780838
                       Trimethoprim                      3.029851
                       Unasyn                            7.000000
                       Uribel                            8.967213
                       Vibramycin                       10.000000
Length: 305, dtype: float64
```

*Fig. 8: groups of diseases and medicines based on ratings and count*

| Drug | Rating_Weighted_Avg | Disease | Symptoms | Review | Sentiment | Rating | UsefulCount |
|---|---|---|---|---|---|---|---|
| Cefixime | 10.0 | urinary tract infection | [' burning_micturition', ' bladder_discomfort'... | "My daughter was born with urinary reflux, the... | Negative | 10 | 7 |
| Doribax | 10.0 | urinary tract infection | [' burning_micturition', ' bladder_discomfort'... | "It is proved itself it is an excellent drug o... | Positive | 10 | 1 |
| Doripenem | 10.0 | urinary tract infection | [' burning_micturition', ' bladder_discomfort'... | "It is proved itself it is an excellent drug o... | Positive | 10 | 1 |
| Lactobacillus acidophilus | 10.0 | urinary tract infection | [' burning_micturition', ' bladder_discomfort'... | "This has been a miracle for me, and it was by... | Negative | 10 | 189 |
| Macrodantin | 10.0 | urinary tract infection | [' burning_micturition', ' bladder_discomfort'... | "Started experiecing pain peeing and went for ... | Positive | 10 | 28 |

*Fig. 9: Final dataset for medicine recommendation*

## ***Results –***

The KNN model gave the most promising results out of the 3 models trained. The accuracy score of 98.98% and F1-score of 99.0 was obtained in case of KNN model. The SVC model did not give those promising results compared to KNN as only 96.88% of accuracy and 96.88 of F1-score was obtained. But using the GridSearchCV the accuracy was improved for the SVC model. Accuracy of 97.02% and F1-score of 97.0 was obtained. Table 1. Shows the comparison of all the 3 models based on their accuracy and F1-score. For providing the recommendations the predicted disease was taken as input and the disease was recommended based on the highest weighted average and the count if the ratings. Fig. 10 shows the manual test case 1 for combined disease prediction and recommendation. Fig. 11 shows the manual test case 2 for combined disease prediction and recommendation.

*Table 1: Comparison table*

| Model | Accuracy | F1-Score |
|---|---|---|
| KNN | 98.98 | 99.0 |
| GridSearchCV | 97.02 | 97.0 |
| SVC | 96.88 | 96.88 |

```
--------------------------------------------------------------------------------
Symptoms:  ['continuous_sneezing', 'scurring', 'skin_peeling', 'silver_like_dusting', 0, 0, 0]
--------------------------------------------------------------------------------
The predicted disease is: Acne
--------------------------------------------------------------------------------
Recommended drugs for Acne are:  ['Aldactone' 'BenzEFoam' 'Benzoyl peroxide / sulfur']
--------------------------------------------------------------------------------
```

*Fig. 10: disease prediction with recommendation [case 1]*

```
--------------------------------------------------------------------------------------------------------------------------------
Symptoms:  ['muscle_wasting', 'yellowish_skin', 'dark_urine', 'throat_irritation', 'drying_and_tingling_lips', 'muscle_pain', 'altered_sensorium']
--------------------------------------------------------------------------------------------------------------------------------
The predicted disease is: Migraine
--------------------------------------------------------------------------
Recommended drugs for Migraine are:  ['Methergine' 'Imitrex Nasal']
--------------------------------------------------------------------------
```

*Fig. 11: disease prediction with recommendation [case 2]*

## *Discussion –*

This paper proposes an AI based solution for a general disease prediction based on symptoms and medicine recommendation by using various machine learning algorithms. The proposed approach is based on 4 main steps 1.) analyzing the symptoms of different diseases 2.) disease prediction 3.) sentiment analysis for the medicine review and calculating the weighted average rating and 4.) medicine recommendation for the predicted disease. Out of the 3 ML models trained KNN was able to give the highest accuracy of 98.98% which was comparatively better to existing model [2]. Although the proposed approach was able to give highest accuracy it was only able to predict only 12 diseases out of the 41 available diseases from the original dataset. This showed that the prediction of the disease largely depends on the dataset in hand. The reliability of the medicine recommendation can be improved in future by including the demographic details of the patients during the training process. The future work of the project can be to collect the appropriate disease-symptom dataset with medicine recommendation data or to formulate a new dataset based on the dataset which was used in this paper with collaboration with the domain experts. Also, various models like CNN, LSTM can also be used for the prediction of the diseases.

## *References –*

[1] Education // News, OCTOBER 7, 2021, “Artificial Intelligence in Medical Diagnosis”

[2] Anjum Unnisa, Kotha Sreni & Ruchika Rachakonda, “SYMPTOM BASED DISEASE PREDICTION AND MEDICINE RECOMMENDATION SYSTEM”, High Technology Letters, Volume 26, Issue 7, 2020, ISSN NO : 1006-6748

[3] Dr. Meera Gandhi, Vishal Kumar Singh, Vivek Kumar, “IntelliDoctor – AI based Medical Assistant “, 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 978-1-7281-1599-3/19/$31.00

[4] Radhika V M, Dr. Swaraj K P, “Movie Genre Prediction and Recommendation Using Deep Visual Features from Movie Trailers”, 2020 International Conference on Power, Instrumentation, Control and Computing (PICC) | 978-1-7281-7590-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/PICC51425.2020.9362496