e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

DISEASE PREDICTION SYSTEM USING DEEP LEARNING

Ayush Sharma*1, Dr. Neha Agrawal*2

*1Dept. Of Information Technology, Maharaja Agrasen Institute Of Technology, New Delhi, Delhi, India.

*2Assistant Professor, Dept. Of Information Technology, Maharaja Agrasen Institute Of Technology, New Delhi, Delhi, India.

ABSTRACT

The paper proposes a model in which the user can enter unstructured signs or pick out the signs and symptoms counseled by using the gadget, primarily based on which, a list of probable sicknesses is provided again to the user. In addition, the user can select any of the output sicknesses to get greater information about its other symptoms, causes, analysis, feasible remedy, and so forth. To help the person higher recognize the ailment and cutting-edge clinical situation. The device also indicates different signs based on those that the consumer has enter. The gadget may be utilized by someone with constrained clinical information as properly comfortably and can come reachable in early sickness detection and analysis. It could also gain customers that are reluctant to go to hospitals on the onset of sweet sixteen signs. This will provide them with a primary idea of the severity of the sickness.

Keywords: Machine Learning, Deep Learning, Naïve Bayes, Support Vector Machine, Classification Algorithms.

I. INTRODUCTION

AI is constantly utilized in classification and relapse issue which helps in abusing the examples in informational collection. It is more precise methodology in anticipating qualities and set the imprint in exact outcomes in earlier years. It has set up as an inventive field to recognize the secret examples in gigantic informational collection. Clinical science is another field where enormous measure of information is created utilizing distinctive clinical reports and other patient manifestations. In this paper, we have utilized calculations like support vector machine, Naive Bays and neural networks which chips away at every indication as for infection. We need learning calculation which will handle high dimensional issues and computational speed. Deep learning has solved mostly problems by giving dense layer and hidden layers concept. With the utilization of the web and all assets accessible to the user, appropriate sicknesses are utilized, and dependent on that legitimate medicine is done which is valuable to all individuals. It is exceptionally useful for specialists and patients to think better about the sickness with no clinical trials or whatever else. The identification of illness dependent on infection is a mind boggling game. Being new to organic terms, the users feed the indications in non-specialized or characteristic terms which add intricacy in anticipating infections. The primary goal is to foster a novel design that could acknowledge and deal with such sort of user inquiries by utilizing strategies like question development utilizing a thesaurus, equivalent coordinating, and indication idea that will permit sickness forecast with more prominent exactness dependent on user input. We have scratched information from the web and produced dataset which can be utilized in future research. Inquiry search recovery and coordinating are utilized in such issues to accomplish expectation.

II. LITERATURE REVIEW

The author has talked about supervised and unsupervised learning following various steps for executing machine learning algorithms. They have used Support Vector Machine, Decision Trees and Naive Bayes and showed that SVM is more accurate than others two. Also decision tree gives inaccurate results of 74%. Here the recommender system is generated where patients logs are updated and past predictions are stored to better predict future results. In the paper “A study on data mining prediction techniques in healthcare sector” [2] the fields that mentioned are, information Discovery method (KDD) is that the method of adjusting the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial removal of implicit, unknown and doubtless helpful data from information in databases. The repetitious method consists of the subsequent steps: information cleansing, information integration, information choice, information transformation, data processing, Pattern analysis, Knowledge. Healthcare data processing prediction supported data processing techniques are as follows: Neural network, Bayesian Classifiers, call tree, Support Vector Machine. The paper states the comparative study of various aid predictions, Study of information mining techniques and tools for

prediction of cardiovascular disease, numerous cancers, and diabetes, disease and medicine conditions. Few limitations are that if attributes are not related then Decision trees prediction is less accurate and

ANN is computationally intensive to train also it does not lead to specific conclusion. Dangare, Chaitrali & Apte, Sulbha. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications. 47. 44-48. 10.5120/7228-0076.

Dangare has introduced a neural network system for heart disease having 12attributes and 2 newly added attributes are taken where data mining techniques like decision trees, naive bayes and Neural Networks are used. This study used confusion matrix to show accuracy of prediction and used structured data in the form of rows and columns. The study showed that neural networks gives accurate results where MLPNN(Multi-layer Perceptron Neural Networks) is used having input layer, hidden layer, output layer is formed. The paper "An approach to devise an Interactive software solution for smart health prediction using data mining" [5] aims in developing a computerized system to check and maintain your health by knowing the symptoms. It has a symptom checker module which actually defines our body structure and gives us liability to select the affected area and checkout the symptoms. Technologies implemented in this paper are: The front end is designed with help of HTML, Java Script and CSS. The back end is designed using MySQL which is used to design the databases. This paper also contains the information of testing like Alpha testing which is done at server side or we can say at the developer's end, this is an actual testing done with potential users or as an independent testing process at server end. And Beta testing is done after performing alpha testing, versions of a system or software known as beta versions are given to a specific audience outside the programming team. Only the limitation of this paper is it suggests only the award-winning doctors and not the nearby doctors to the patient.

III. DATASET USED

The already accessible dataset is confined to a specific piece of human body infection and is likewise more modest in volume. Subsequently, the dataset of infection and their symptoms has been scratched from the web by running the Python script. The dataset comprises of illnesses and their indications, which are gotten from the accompanying sources: Diseases: The rundown of sicknesses has been recovered from the National Health Portal of India (https://www.nhp.gov.in/infection a-z), created and kept up by Center for Health Informatics (CHI). The content gets the HTML code of the page and concentrates the infection list by sifting values in HTML labels. Symptoms: The content uses the Google Search bundle to perform looking and bring the illness' Wikipedia page among the different indexed lists acquired. The HTML code of the page is handled to bring the indications of the infection utilizing the 'infobox' accessible on the Wikipedia page. Figure 1 shows an illustration of Wikipedia's infobox. Every one of the indications are extracted and a word reference is made with key as disease and symptoms as value. Further, every disease is treated as the name and all symptoms are treated as explicit qualities or segments. It shows the orderly progression of steps engaged with information scratching. The scratching script brings more than 261 unique infections that structure the mark and 500+ indications. The manifestations are then pre-prepared to eliminate comparative symptoms with various names (For ex-abundant, migraine and agony in the brow). This is finished by discovering the equivalent words for every manifestation and processing Jaccard Coefficient for sets of indications. In the event that the score is more noteworthy than the sift old, both the manifestations are basically the same and one of them can be eliminated.

Fig 1: Wikipedia Infobox

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

in the event the Jaccard(Symptom1,Symptom2)>threshold:
Symptom2->Symptom1

To increase the dataset, every illness' symptoms are gotten, blends of the manifestations are made and added as new columns in the dataset. For instance, an illness A, having 5 symptoms, presently has a total of (2^5- 1) passages in the dataset. The dataset, subsequent to preprocessing and augmentation, contains around 8835 columns with 489 remarkable symptoms.

Proposed Solution Sketch

On a general note, the system prompts the user to enter symptoms based on which model predicts diseases with the highest probability and scores. Figure 1 describes the process of disease prediction from user input symptoms. The following subsections discuss each module in detail.

IV. SYMPTOM PREPROCESSING

The system accepts symptom(s) in a single line, separated by comma(,). Subsequently, the follow-ing preprocessing steps are involved:

- Split symptoms into a list based on comma.
- Convert the symptoms into lowercase
- Removal of stopwords
- Tokenization of symptoms to remove any punctuation marks
- Lemmitization of tokens in the symptoms

Symptom Expansion using Synonyms

Each symptom is expanded by appending a list of its synonyms. The synonyms are taken from thesauras.com (https://www.thesaurus.com/) and Princeton University's WordNET (https:// wordnet.princeton.edu/) available in Python. Each symptom is broken into its combinations for finding the synonyms set.

```
Common co-occuring symptoms:
0 : headache
1 : testicular pain
2 : vomiting
3 : barky cough
4 : sore throat
Do you have have of these symptoms? If Yes, enter the indices (space-separated), 'no' to stop, '-1' to skip:
no


    Top 10 disease based on Cosine Similarity Matching :
)
0. Disease : Brucellosis         Score : 0.62
1. Disease : Influenza   Score : 0.33
2. Disease : Perennial Allergic Conjunctivitis   Score : 0.32
3. Disease : Asthma      Score : 0.29
4. Disease : Rocky Mountain spotted fever        Score : 0.1
5. Disease : Leptospirosis       Score : 0.07
6. Disease : Rift Valley fever   Score : 0.07
7. Disease : Malaria     Score : 0.06
8. Disease : Black Death         Score : 0.06
9. Disease : Shigellosis         Score : 0.06

More details about the disease? Enter index of disease or '-1' to discontinue and close the system:
-1
```

Figure 2

```
Top 10 diseases predicted based on TF_IDF Matching :

0. Disease : Influenza    Score : 5.75
1. Disease : Perennial Allergic Conjunctivitis    Score : 5.56
2. Disease : Asthma       Score : 4.47
3. Disease : Brucellosis          Score : 4.47
4. Disease : Acute encephalitis syndrome          Score : 1.29
5. Disease : Anthrax      Score : 1.29
6. Disease : Aseptic meningitis           Score : 1.29
7. Disease : Black Death          Score : 1.29
8. Disease : Bubonic plague       Score : 1.29
9. Disease : Chagas disease       Score : 1.29

More details about the disease? Enter index of disease or '-1' to discontinue:
0

Influenza
Other names -  Flu, the flu, Grippe
Specialty -  Infectious disease
Symptoms -  Fever, runny nose, sore throat, muscle pain, headache, coughing, fatigue
Usual onset -  1–4 days after exposure
Duration -  2–8 days
Causes -  Influenza viruses
Prevention -  Hand washing, flu vaccines
Medication -  Antiviral drugs such as oseltamivir
Frequency -  3–5 million severe cases per year
Deaths -  Up to 650,000 deaths per year
```

Figure 3

Higher cosine similarity represents a higher similarity between the disease and the query vector. The scores are sorted based on a decreasing score and a list of top K diseases is obtained. Figure 2 shows the predicted list of diseases with a cosine similarity score. Figure 3 shows the information about the diseases collected from the wikipedia infobox so the user needs the proper knowledge about the diseases.

Disease Detail

The user can select any of the disease output by the model and view the details of that disease in the console.

V. RESULT

Evaluation of the dataset is done by applying various machine learning algorithms and comparing the accuracy obtained from them. Here Neural network shows highest I.e 91.22% accuracy while multinomial naive bayes and support vector machine shows 84% and 87% accuracy.Figure 4 shows the comparison between neural network, naive bayes and support vector machine.

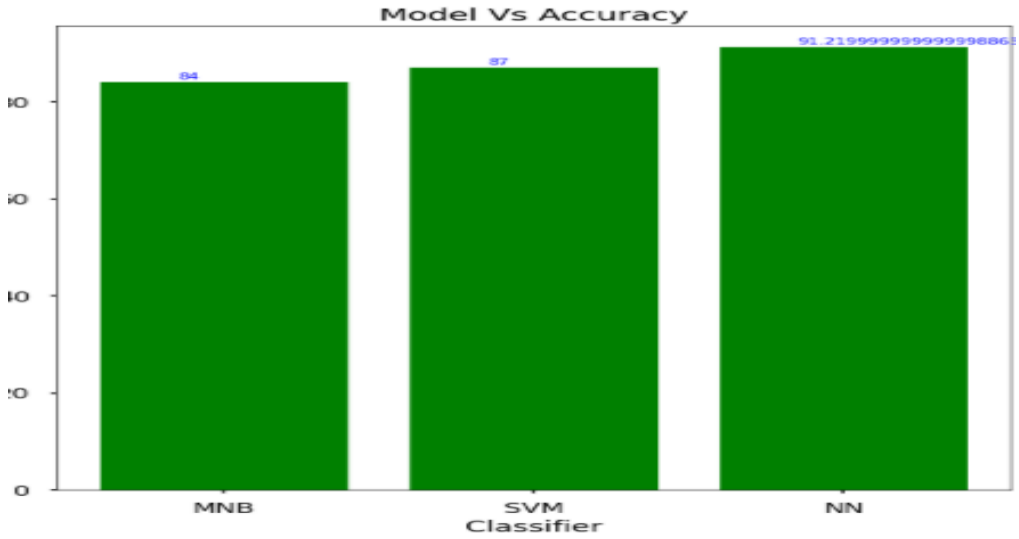

Figure 4

VI. CONCLUSION

In this paper, we concluded that Neural Network (Deep Learning) is better for our dataset in comparison to other machine learning algorithms. It checks all the probabilities of each feature independent of each other. We have seen that this system have lot more work to do in near future as It was able to predict diseases based on three symptoms. We can use more symptoms to predict diseases and thus make it more accurate. Also we can give the probability of having different disease. Thus if any critical case happens, then we can refer the case to best possible hospital early as possible. Specialists can use it for their support and it can also stand out as an alternative solution in localities where medical help is far away.

VII. REFERENCE

[1] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The 'Big Data' Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Office, 2016.

[2] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning over Big Dat a From Healthcare Communities," in IEEE Access, v ol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACC ESS.2017.2694446.

[3] Shubham Bind et al, / (IJCSIT) International Journal of Computer Science and Information Technologies ,Vol. 6 (2) , 2015, 1648-1655

[4] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Le arning Techniques," in IEEE Access, vol. 7, pp. 81542- 81554, 2019

[5] Md. Martuza Ahamad, Sakifa Aktar, Md. Rashed-Al-Mahfuz, Shahadat Uddin, Pietro Liò, Haomi ng Xu, M atthew A. Summers, Julian M.W. Quinn, Mohammad Ali Moni, "A machine learning model to identify ear ly stage symptoms of SARS- Cov-2 infected patients", Expert Systems with Applications, Volume 160,20 20,113661.

[6] Podgorelec, V., Kokol, P., Stiglic, B. et al. Decision Trees: An Overview and Their Use in Medicine. Journal of Medical Systems 26, 445–463 (2002).

[7] Dangare, Chaitrali & Apte, Sulbha. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications. 47. 44-48. 10.5120/7228-0076.

[8] G.M. Cramer, R.A. Ford, R.L. Hall,"Estimation of toxic hazard—A decision tree approach, Food and Cosmetics Toxicology", Volume 16, Issue 3,1976,Pages 255-276,ISSN 0015-6264.

[9] Shadab Adam Pattekari and Asma Parveen "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, Pages 290-294.