Grenze International Journal of Engineering and Technology, Jan Issue

An Automated Disease Prediction using AI

Honey Jain[1], Rohan Benhal[2] and Tanmayee Parbat[3]
[2, 3]University of California, Riverside
Email: tanmayee.parbat01@gmail.com
[1]Drexel University

***Abstract*—Data Mining and Machine Learning plays most motivating space of exploration that become generally well known in wellbeing association. It likewise has a crucial impact to reveal new examples in therapeutic science and administrations affiliation which subsequently obliging for every one of the gatherings related with this field. This undertaking expect to frame a symptomatic model of the normal sicknesses dependent on the manifestations by utilizing information mining method like arrangement in wellbeing space. In this paper, we will utilize AI calculations and profound realizing which can be used for health care diagnosis. In this paper we are proposing the disease identification using symptoms and images of malaria cells. Chest x-ray images of covid and pneumonia disease. Accordingly we are recommending the precautions and hospital to the patient.**

***Index Terms*— Disease Prediction, Deep Learning and Neural Networks, Support Vector Machine, KNN.**

I. INTRODUCTION

Human is master in getting data, while machine is master at communicating and handling information. In this paper, we propose a model for patient side effect closeness examination by exploiting the machine's capacity to process information. The model utilized patient's portrayals of indications to remove key data and accomplish early expectation and mediation. Consequently, the precision of likeness examination model to a great extent decides the adequacy of infection expectation. Accurately predicting diseases plays a significant role in public health, especially at the early stage which allows patients to take prevention treatments in time. With the growing volume and availability of electronic health records (EHRs), predictive modeling tasks for disease progression and analysis have obtained increasing interest from researchers. The EHR data are temporally sequenced by patient visits with each visit represented as a set of high dimensional clinical events. Mining EHRs is especially challenging compared to standard data mining tasks, due to its noisy, irregular and heterogeneous nature. Recently, deep learning and machine learning approaches have been widely adopted and rapidly developed in patient representation learning.

II. LITERATURE SURVEY

Qiuling Suo, Fenglong Ma et al. [1] Stated that presenting a novel time fusion CNN framework to simultaneously learn patient representations and measure pairwise similarity. Compared to a traditional CNN, our time fusion CNN can learn not only the local temporal relationships but also the contributions from each time interval. Along with the similarity learning process, the output information which is the probability

Grenze ID: 01.GIJET.9.1.314

distribution is used to rank similar patients.

Prabakaran.N and Kannadasan et al. [2] compares recent healthcare data against data from that particular baseline distribution and hence classifies subgroups of the given data. In addition, the data sample data used is first tested against many types of classifiers and various other proposed test scores have been evaluated.
Peiying zhang, Xingzhe huang et al. [3] proposed a sentence similarity model to carry out symptom similarity analysis to achieve elementary disease prediction and early intervention, which makes use of word embedding and convolutional neural network (CNN) to extract a sentence vector that contains keyword information about the patient's feelings and symptoms. In order to increase the accuracy of sentence similarity computation, this model integrated syntactic tree and neural network into the computation process. Our main innovation is to use symptom similarity analysis model for disease prediction and early intervention. In addition, the SPO kernel is also one of the innovations.
Zhaoqian Lan. Guopeng Zhou [4] proposes AI-assisted prediction system, which leverages data mining methods to reveal the relationship between the regular physical examination records and the potential health risk. It can predict examinees' risk of physical status next year based on the physical examination records this year. The system provides a user-friendly interface for examinees and doctors. Examinees can know their potential health risks while doctors can get a set of examinees with potential risk.
Pär Salander et al. [5] proposed that Most spouses witnessed months of global dysfunction preceding the symptom leading to physician consultation. The patient factors 'less alien symptoms', 'personality change' and 'avoidance'; the spouse factors 'spouse's passivity' and 'spouse's successive adaptation'; and the physician factors 'reasonable alternative diagnosis', 'physician's inflexibility' and 'physician's personal values' were identified as obstacles on the pathway to appropriate medical care.

III. Problem Statement

1] Detection of covid, malaria and pneumonia lesions using image data and symptoms with more two more diseases such as Heart disease and diabetes.
2] It is a tool which will take input as the CT scan images and it will predict the possibilities of the disease and its stages using deep learning and disease prediction based on symptoms using machine learning.
3] To give higher exactness over past research.

IV. Proposed Method and Algorithm

A. Proposed Methodology

In a propose system, we are proposed experiment on detecting disease like Diabetes, malaria, Migraine, Heart Failure and Covid and recommend hospitals for specific disease with limited set of supervised data.
We come through a wide range of different and major algorithms for predicting the monotonous diseases with comprehensible symptoms while working in the field of Supervised Machine Learning such as support vector machine. We are recommending the nearest hospitals from the patients location through KNN. By using CNN we are classifying the disease through images.

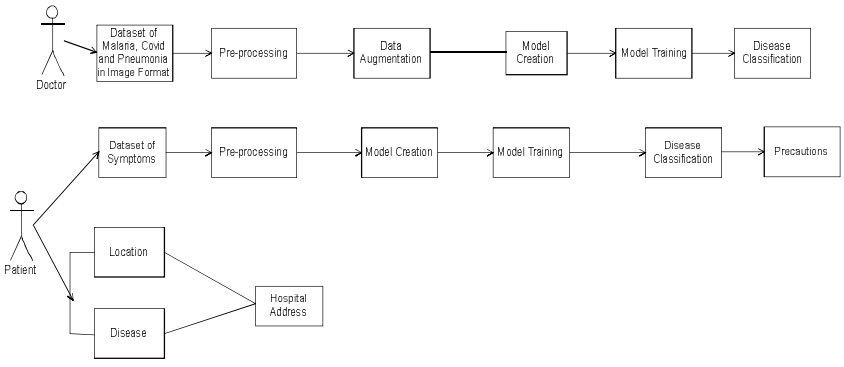

Fig 1. Proposed Architecture

B. Dataset

In this project we are collecting the data from kaggle platform. We require two different datasets one for symptoms in Table 1 and another for disease prediction based on image as shown in figure2.

TABLE I. SYMPTOMS

Pneumonia	Covid19
cold	Dry-Cough
greenish cough	Sore-Throat
yellow cough	Difficulty-in-Breathing
cough with blood	Tiredness
Fever	Fever
sweating	Body Pain
shaking	
shortness of breath	
Rapid breathing	
shallow breathing	
chest pain	
low energy	
Loss of appetite	
fatigue	
Nausea	
vomiting	

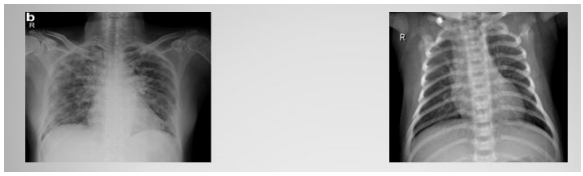

Figure2. Chest X-Ray Images

C. Algorithms

Tfidf

Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF). The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents. Document frequency is the number of documents containing a specific term. Document frequency indicates how common the term is. Inverse document frequency (IDF) is the weight of a term, it aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents.

Convolutional Neural Networks(CNN)

Convolutional Neural Networks (which are additionally called CNN/ConvNets) are a kind of Artificial Neural Networks that are known to be tremendously strong in the field of distinguishing proof just as picture order.
Four main operations in the Convolutional Neural Networks are shown as follows:

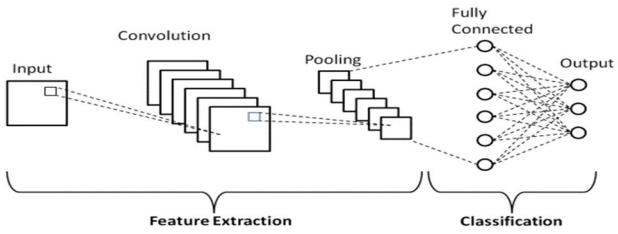

Figure 3. Architecture of CNN

(i) Convolution

The principle utilization of the Convolution activity if there should be an occurrence of a CNN is to recognize fitting highlights from the picture which goes about as a contribution to the primary layer. Convolution keeps up the spatial interrelation of the pixels This is finished by fulfillment of picture highlights utilizing miniscule squares of the picture. Convolution equation. Every picture is seen as a network of pixels, each having its own worth. Pixel is the littlest unit in this picture grid. Allow us to take a 5 by 5(5*5) framework whose qualities are just in twofold (for example 0 or 1), for better agreement. It is to be noticed that pictures are by and large RGB with upsides of the pixels going from 0 - 255 i.e 256 pixels.

(ii). ReLU

ReLU follows up on a rudimentary level. All in all, it is an activity which is applied per pixel and overrides every one of the non-positive upsides of every pixel in the component map by nothing.

(iii). Pooling or sub-sampling

Spatial Pooling which is likewise called sub-sampling or down sampling helps in lessening the elements of each element map yet even at the same time, holds the most

important data of the guide. Subsequent to pooling is done, in the long run our 3D element map is changed over to one dimensional component vector.

SVM

Generally the use of machine learning algorithm is used for the recommendation. In this paper, the support vector machine will suggest the disease. SVM is a supervised machine learning algorithm which works based on the concept of decision planes that defines decision boundaries. A decision boundary separates the objects of one class from the object of another class. Support vectors are the data points which are nearest to the hyper-plane. Kernel function is used to separate non-linear data by transforming input to a higher dimensional space.

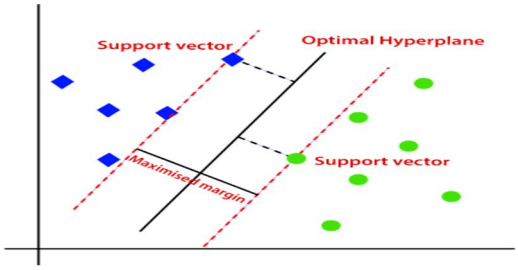

Fig 4. SVM Architecture

KNN

A refinement of the k-NN characterization calculation is to gauge the commitment of every one of the k neighbors as indicated by their distance to the inquiry point, giving more prominent load to nearer neighbors. The KNN classifier recommending the emergency clinic subtleties for infection dependent on the closest distance.

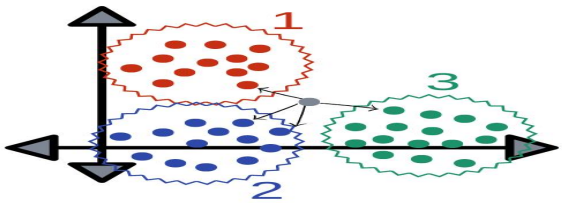

Fig 5. KNN Architecture

$d=\sqrt{((x2-x1)^2+(y2-y1)^2)}$

d=distance

x1, x2, y1, y2 = data points

V. Results & Discussion

In our experimental setup, as shown in table 2, the total numbers of 800 of trained images for three diseases such as malaria, covid and pneumonia and 130 new images were tested. These images go through CNN framework by following feature extraction using our image processing module. Then our trained model of classification of diseases get classifies the image into specifies disease. We get the accuracy 89.45% at 50 epochs.

TABLE II. CLASSIFICATION OF DATA

Sr. No.	Category	Number of Images
1	Training	800
2	Testing	130

VI. CONCLUSION

In the paper, we are invented Symptoms based disease prediction is based on Machine Learning and disease prediction by image data using deep learning. This system is useful for both doctors and patient. A patient can predict the disease based on the symptoms and recommend the hospital also and doctors will identify the disease uploading chest x-ray images. Our system takes symptoms and image as input and gives output as disease, possible precautions.

REFERENCES

[1] Qiuling Suo∗, Fenglong Ma et al. Personalized Disease Prediction Using a CNN-Based Similarity Learning Method 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

[2] Prabakaran.N and Kannadasan.R "Prediction of Cardiac Disease Based on Patient's Symptoms "Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2.

[3] PEIYING ZHANG , XINGZHE HUANG , AND MAOZHEN LI "Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis" Engineering and Design, Brunel University London, Uxbridge UB8 3PH, U.K.

[4] Qiuling Suo∗, Fenglong Ma et al. AI-assisted Prediction on Potential Health Risks with Regular Physical Examination Records Ieee transactions on knowledge and data science 19 july.

[5] Arief Setyanto, Hartatik and Mohammad Badri Tamam **"**Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms**"** 020 2nd International Conference on Cybernetics and Intelligent System (ICORIS).