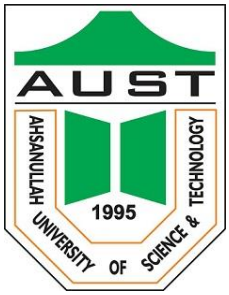

AHSANULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

DISEASE PREDICTION BASED ON SYMPTOMS USING NEURAL NETWORK

Course No: CSE 4238

Course Name: Soft Computing Lab

Group Members

Name: Md. Robiul Islam **ID:** 16.02.04.080

Name: Mir Hasibul Hasan **ID:** 16.02.04.081

Name: Md. Yeasir Arafat **ID:** 16.02.04.093

1. Introduction

An optimum strategy is used to discover an optimal sequence of diseases in order to predict the disease once the patient experiences symptoms in an efficient manner and to improve the execution performance of the DPS (Disease Prediction System) in terms of accuracy. Because ANN (Artificial Neural network) produces high accuracy prediction levels, it is used to implement a prediction system. To predict the disease, an artificial neural network is prepared. The accuracy in disease prediction is compared to that of other approaches. **Example: let a person having some symptoms of a disease but he can't identify it then if his symptoms are given input into our project he will get a specification that he is suffering from 'X' disease.**

a. **Motivation:**

   The widespread adoption of smartphones creates an enormous potential to improve healthcare services. Numerous apps, sensors, and devices are developed for health self-management purposes. So, what is our motivation for this project? Well, the quality of healthcare services in Bangladesh has led to the loss of faith in healthcare providers. There is low utilization of public health facilities. The rate of wrong treatment in Bangladesh is so high. Moreover, many people die over the wrong treatment. This is an unacceptable case. Especially the people of our country who live in rural areas sometimes can't decide where to go which category of doctor they would face. Then they take their own decision to go for a category to select a doctor based on some traditional parameters. This traditional culture delays their proper treatment; sometimes it drives them to death. This is why we're here to introduce our project.

b. The task is challenging because we can't get real time actual data. So, we used dummy data to train our model. But to specify disease for a particular area or country the nature of data might be different .So, our prediction model is not real stic or applicable for real life. We got more challenges to merge our symptoms, removing redundant data .

2. Related Works

a. We searched for some medical apps running in bangladesh but didn't find anyone to do the same task like our project . As if medical field is a very sensitive issue besdies there are so many rules to store patients' personal database and use it to train (our assumption) so nobody dare to launch it as a real life project. List of some medical apps in Bangladesh:They are PlexusD, Tonic, Praava Health, Doctor Dekhau, SeekMed, Doctorola.com, Doctorbd.com, Daktarachen.com.

b. **(i)** We didn't find any question base system to find out the category of disease
(ii) One of the existing systems PlexusD is taking fees just for getting permission to appoint a doctor, chat with a doctor, and other simple services.

(iii) There is no disease detection system. Patients can't cross-check their disease by their existing medical data.

(iv) None of them use the patient's existing data for future prediction.

(v) None of them can store a patient's unique data so that the doctor can understand what should do and what should be avoided for a specific patient. For example, a patient has a drug allergy but he/she can't memorize it. If the doctor wouldn't know it and the same drug can be suggested. As a result, the patient won't get proper treatment.

c.

(i) Patients can cross-check their disease on their own. When they will have available medical data at their hand then they will input it to our system and get a prediction of being cachectic of a particular disease.

3.Project Objective:

The general disease prediction system predicts chance of presence of a disease present in a patient on the basis of their symptoms. It will also recommend necessary precautionary measures required to treat the predicted disease. The system will initially be fed data from a dataset, the data will then be pre-processed before further process is carried out, this is done so as to get clean data from the raw initial data, as the raw data would be noisy, or flawed. This data will be processed using Data mining algorithms, the system, will be trained so as to predict the disease based on the input data given by the user.

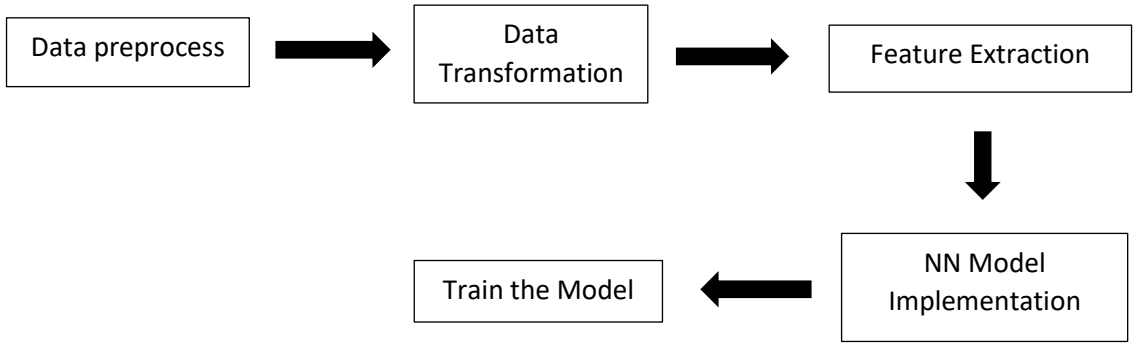

Fig 1: Subtask of disease prediction system

Inputs and Outputs:

User Input	Predicted Output
vomiting weight_loss high_fever yellowish_skin dark_urine abdominal_pain	Jaundice
itching skin_rash fatigue headache red_spots_over_body high_fever loss_of_appetite	Chicken pox
sweating vomiting breathlessness	Heart attack

4.Methodologies:

Our system can identify disease based on given disease symptoms. To make this system we need a dataset which is downloaded from https://www.kaggle.com/itachi9604/disease-symptom-description-dataset?select=dataset.csv this link. Then we pre process the dataset to remove redundant space. After that we make a Bag of Words binary vocabulary based on the symptoms. We found 132 unique symptoms in that dataset. And also making a vocabulary of total number of diseases in that dataset. There were 41 different Disease in that dataset. Then we split the whole dataset into Train and Test as a ratio of 9:1. After that we make our Neural Network model and train it with train dataset and test it with test dataset. Finally, we can predict disease using our train model which takes an input of a list of disease symptoms of a patients.

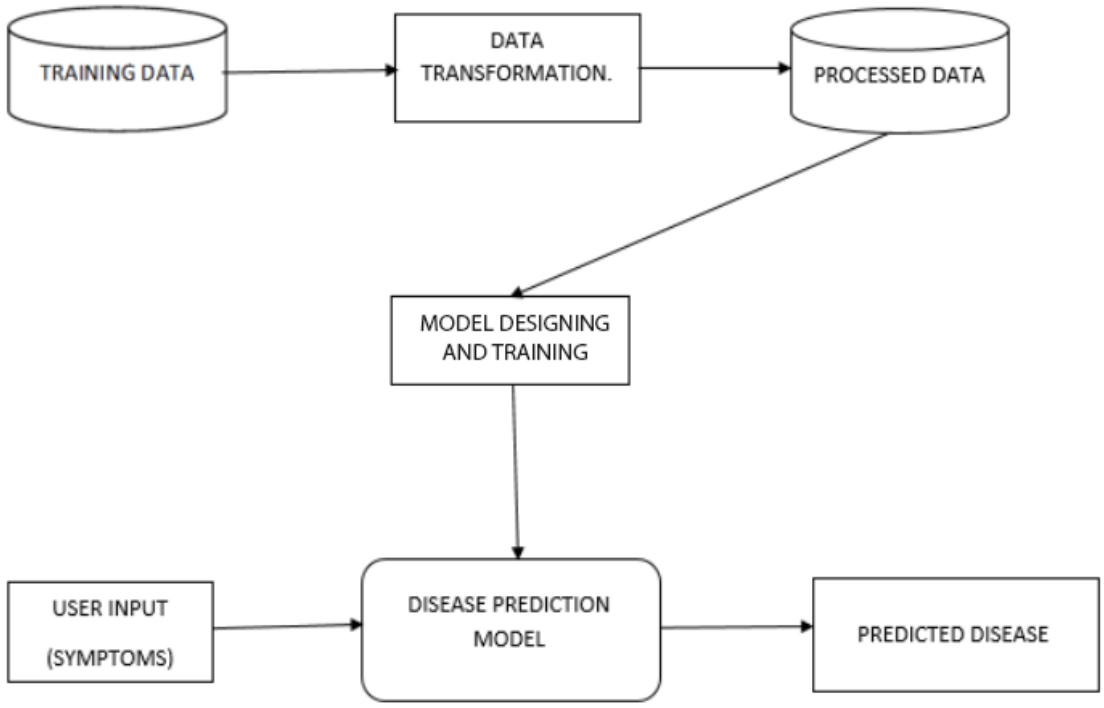

Fig 2: Block Diagram for General disease prediction system.

Neural Network Model:

```
BagOfWordsClassifier(
  (linear_1): Linear(in_features=132, out_features=100,
bias=True)
  (relu_1): ReLU()
  (linear_2): Linear(in_features=100, out_features=64,
bias=True)
  (relu_2): ReLU()
  (linear_out): Linear(in_features=64, out_features=41,
bias=True)
)
```

5.Experiments:

Dataset Statistics:

```
Number of Features in an Input Vector:  132
Number of Classes (Outputs):  41
Number of Vocabulary:  132
Number of Train Samples:  4428
Number of Test Samples:  492
Train and Test split ratio: 9:1
```

Each Class with their number of Train Samples taken for train this model:

Index	Class Labels	Number of Train Samples
0	(vertigo) Paroymsal Positional Vertigo	107
1	AIDS	109
2	Acne	108
3	Alcoholic hepatitis	107
4	Allergy	107
5	Arthritis	106
6	Bronchial Asthma	112
7	Cervical spondylosis	109
8	Chicken pox	102
9	Chronic cholestasis	107
10	Common Cold	106
11	Dengue	106
12	Diabetes	107
13	Dimorphic hemmorhoids(piles)	99
14	Drug Reaction	101
15	Fungal infection	110
16	GERD	107
17	Gastroenteritis	105
18	Heart attack	111
19	Hepatitis B	110
20	Hepatitis C	116
21	Hepatitis D	108
22	Hepatitis E	110
23	Hypertension	103
24	Hyperthyroidism	105
25	Hypoglycemia	111
26	Hypothyroidism	108
27	Impetigo	108
28	Jaundice	109
29	Malaria	106
30	Migraine	110
31	Osteoarthristis	106
32	Paralysis (brain hemorrhage)	112
33	Peptic ulcer diseae	112
34	Pneumonia	110
35	Psoriasis	104
36	Tuberculosis	111
37	Typhoid	106
38	Urinary tract infection	113
39	Varicose veins	110
40	hepatitis A	114

Features with their Indexes:

(0, 'abdominal_pain'),
(1, 'abnormal_menstruation'),
(2, 'acidity'),
'lack_of_concentration'),
(3, 'acute_liver_failure'),
(4, 'altered_sensorium'),
(5, 'anxiety'), (6, 'back_pain'),
(7, 'belly_pain'), (8, 'blackheads'),
(9, 'bladder_discomfort'),
(10, 'blister'),
(11, 'blood_in_sputum'),
(12, 'bloody_stool'),
(13, 'blurred_and_distorted_vision'),
(14, 'breathlessness'),
(15, 'brittle_nails'),
(16, 'bruising'),
(17, 'burning_micturition'),
(18, 'chest_pain'),
(19, 'chills'),
(20, 'cold_hands_and_feets'),
(21, 'coma'),
(22, 'congestion'),
'pain_during_bowel_move'),
(23, 'constipation'),
(24, 'continuous_feel_of_urine'),
(25, 'continuous_sneezing'),
(26, 'cough'),
(27, 'cramps'),
(28, 'dark_urine'),
(29, 'dehydration'),
(30, 'depression'),
'prominent_veins_on_calf'),
(31, 'diarrhoea'),
(32, 'dischromic_patches'),
(33, 'distention_of_abdomen'),
'receiving_blood_transfu'),
(34, 'dizziness'),
(35, 'drying_and_tingling_lips'),
(36, 'enlarged_thyroid'),
(37, 'excessive_hunger'),
(38, 'extra_marital_contacts'),
(39, 'family_history'),
(40, 'fast_heart_rate'),
(41, 'fatigue'),
(42, 'fluid_overload'),
(43, 'foul_smell_of_urine'),
(44, 'headache'),
(45, 'high_fever'),
(46, 'hip_joint_pain'),
(47, 'history_of_alcohol_consumption'),
(48, 'increased_appetite'),
(49, 'indigestion'),
(50, 'inflammatory_nails'),
(51, 'internal_itching'),
(52, 'irregular_sugar_level'),
(53, 'irritability'),
(54, 'irritation_in_anus'),
(55, 'itching'),
(56, 'joint_pain'),
(57, 'knee_pain'),
(58,

(59, 'lethargy'),
(60, 'loss_of_appetite'),
(61, 'loss_of_balance'),
(62, 'loss_of_smell'),
(63, 'malaise'),
(64, 'mild_fever'),
(65, 'mood_swings'),
(66, 'movement_stiffness'),
(67, 'mucoid_sputum'),
(68, 'muscle_pain'),
(69, 'muscle_wasting'),
(70, 'muscle_weakness'),
(71, 'nausea'),
(72, 'neck_pain'),
(73, 'nodal_skin_eruptions'),
(74, 'obesity'),
(75, 'pain_behind_the_eyes'),
(76,

(77, 'pain_in_anal_region'),
(78, 'painful_walking'),
(79, 'palpitations'),
(80, 'passage_of_gases'),
(81, 'patches_in_throat'),
(82, 'phlegm'),
(83, 'polyuria'),
(84,

(85, 'puffy_face_and_eyes'),
(86, 'pus_filled_pimples'),
(87,

(88, 'receiving_unsterile),
(89, 'red_sore_around_nose'),
(90, 'red_spots_over_body'),
(91, 'redness_of_eyes'),
(92, 'restlessness'),
(93, 'runny_nose'),
(94, 'rusty_sputum'),
(95, 'scurring'),
(96, 'shivering'),
(97, 'silver_like_dusting'),
(98, 'sinus_pressure'),
(99, 'skin_peeling'),
(100, 'skin_rash'),
(101, 'slurred_speech'),
(102,'small_dents_in_nails'),
(103, 'spinning_movements'),
(104, 'spotting_urination'),
(105, 'stiff_neck'),
(106, 'stomach_bleeding'),
(107, 'stomach_pain'),
(108, 'sunken_eyes'),
(109, 'sweating'),

```
(110, 'swelled_lymph_nodes'),
(111, 'swelling_joints'),
(112, 'swelling_of_stomach'),
(113, 'swollen_blood_vessels'),
(114, 'swollen_extremeties'),
(115, 'swollen_legs'),
(116, 'throat_irritation'),
(117, 'toxic_look_'),
(118, 'typhos'),
(119, 'ulcers_on_tongue'),
(120, 'unsteadiness'),
(121, 'visual_disturbances'),
(122, 'vomiting'),
(123, 'watering_from_eyes'),
(124, 'weakness_in_limbs'),
(125, 'weakness_of_body'),
(126, 'weight_gain'),
(127, 'weight_loss'),
(128, 'yellow_crust_ooze'),
(129, 'yellow_urine'),
(130, 'yellowing_of_eyes'),
(131, 'yellowish_skin')
```

Sample of a dataset:

```
(tensor([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 1]]), 20)
```

In the above input, these following index values are One:

Index	**Index Label**
*****	**********
39	family_history
41	fatigue
60	loss_of_appetite
71	nausea
130	yellowing_of_eyes
131	yellowish_skin

Label: Hepatitis C

Evaluation Matrix:

The three main metrics used to **evaluate** a classification **model** are accuracy, precision, and recall. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions We used Accuracy to evaluate our proposed system.

Results:

We designed different Neural Network model with same hyperparameter and analyze which model gives us more accurate result. Finally, the above Neural Network Model gives us more accurate result. Then we set different hyperparameters in the same Neural Network model to analyze which gives us more accuracy. Among them we are showing best three hyperparameter results in the following section.

Best three hyperparameters:

	Setting 1	*Setting 2*	*Setting 3*
Num of Iterations	20,000	40,000	30,000
Batch Size	20	20	20
Learning Rate	0.01	0.001	0.1

Accuracy Table for different hyperparameters:

ITERATION	SETTING 1	SETTING 2	SETTING 3
500	**13.01**	**0.69**	**14.02**
1,000	**41.86**	**2.357**	**30.08**
1,500	**50.40**	**6.09**	**38.61**
2,000	**53.05**	**12.19**	**37.81**
2,500	**67.07**	**16.05**	**63.01**
3,000	**97.96**	**16.66**	**88.82**
3,500	**100.0**	**22.15**	**100.0**
4,000	**100.0**	**28.86**	**99.79**
4,500	**100.0**	**31.09**	**100.0**
5,000	**100.0**	**32.52**	**100.0**

Loss Graphs for those settings:

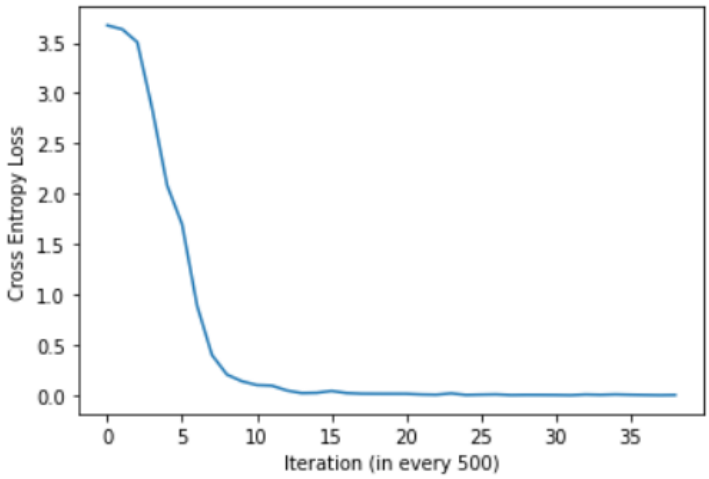

Fig 3: Loss graph for setting 1

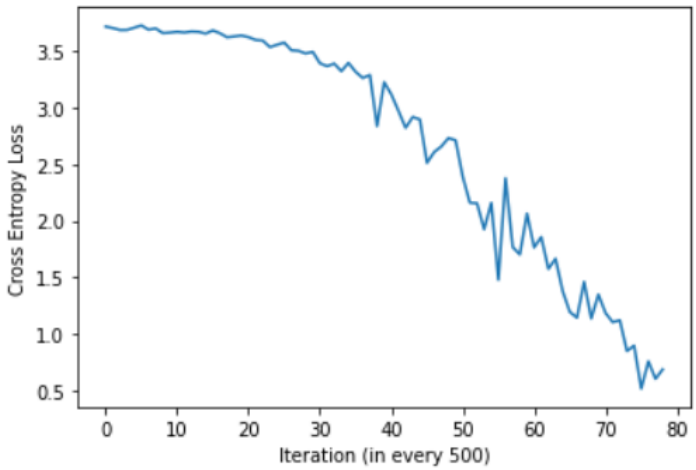

Fig 4: Loss graph for setting 2

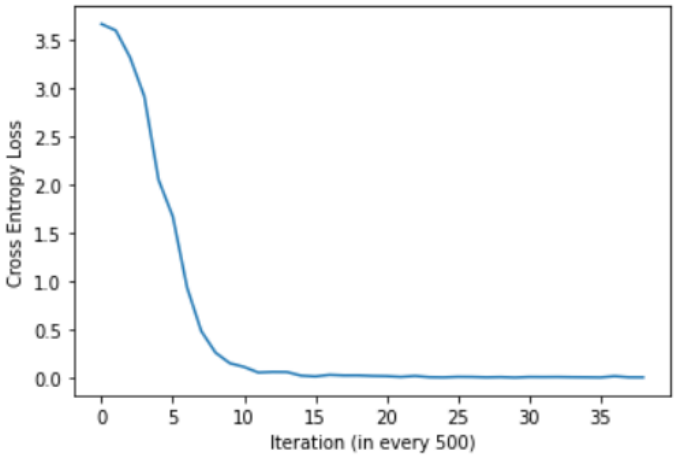

Fig 5: Loss graph for setting 3

From the above data we can say too small learning rate takes more time to fit the model, Setting 1 and setting 3 gives us 100% accuracy. That two hyperparameter is good for our model. We can use any of them to train our disease prediction model.

6.Conclusion:

We have tried our best to give an overview of our project. The system provides a solution to some lacking that we have found on similar systems. We believe if anybody launch our project as a real life application their system will give users more facilities more flexibility rather than any other existing system.