BOHR International Journal of Computer Science
2022, Vol. 1, No. 1, pp. 57–60
DOI: 10.54646/bijcs.2022.10
www.bohrpub.com

METHODS

Disease prediction using machine learning

G. Vasu Sena, K. Rajinikanth, Mohammed Khaja Faizan* and D. Rohit Rajan

Student, Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

***Correspondence:**
Mohammed Khaja Faizan,
fabulousfaizan1234@gmail.com

Received: 12 May 2022; **Accepted:** 01 July 2022; **Published:** 14 July 2022

Predicting disease at an early stage becomes critical, and the most difficult challenge is to predict it correctly along with the sickness. The prediction happens based on the symptoms of an individual. The model presented can work like a digital doctor for disease prediction, which helps to timely diagnose the disease and can be efficient for the person to take immediate measures. The model is much more accurate in the prediction of potential ailments. The work was tested with four machine learning algorithms and got the best accuracy with Random Forest.

Keywords: machine learning, random forest, disease prediction, Naive Bayes

Introduction

The main goal of our project is to provide the disease name by taking the symptoms from the user or patients. Nowadays everything is available on the internet, so we thought of predicting the disease based on the symptoms that are given by the customer online. It is an interactive system that takes symptoms from the customer. The customer has to provide a minimum of 2 symptoms that they are suffering from.

The system responds effectively graphical user interface (GUI) to make it look like or feels like it is a live interaction. You can create this type of disease prediction using machine learning algorithms as well as artificial algorithms to enquire, identify, and respond to the customer.

Random forest algorithm

(1) Random forest selects k number of records randomly from data having m records.
(2) A separate decision tree is created for each sample.
(3) Output is produced from every decision tree.
(4) The result is based on averaging for classification and regression. Random forest is considered one of the effective algorithms used in classification.

Decision tree algorithm

Decision trees are commonly employed for classification. A decision tree is a classifier with a tree structure in which features are represented by internal nodes and the branches of the tree represent decision rules. The decision tree has two nodes. The judgment or test is made based on the dataset's properties.

Naive Bayes algorithm

The algorithm that is used in the classification of binary and multiclass is the Naive Bayes algorithm. The Naive Bayes algorithm is very simple and easy to understand, and the Naive Bayes algorithm provides good output for a wide range of output P (class1| data1) = (P (datallclassl) × P (classl))/P (datal). With the help of the Naive Bayes algorithm, we can calculate the probability of a piece of data belonging to a given class.

K-nearest neighbor (KNN)

A pattern has been found to link the data and results, which helps in improving the recognition with each iteration. It involves the following steps:

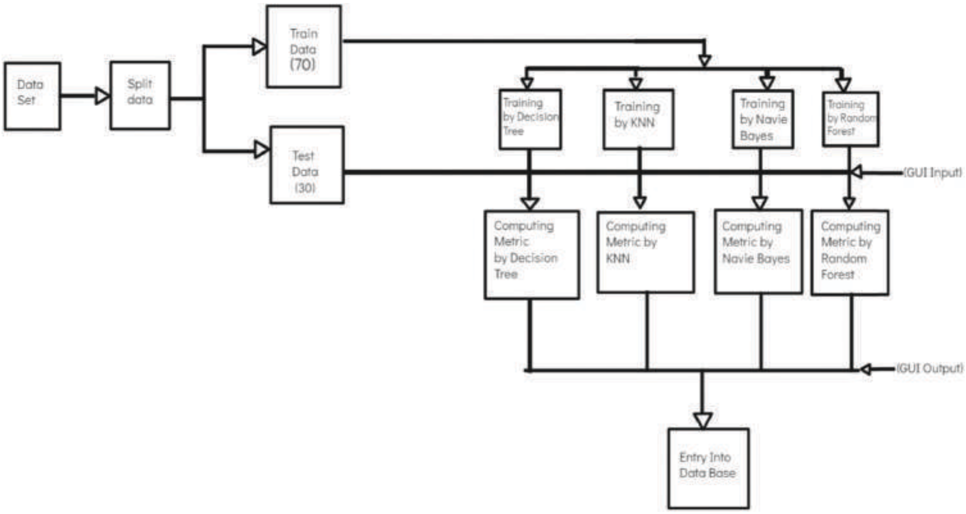
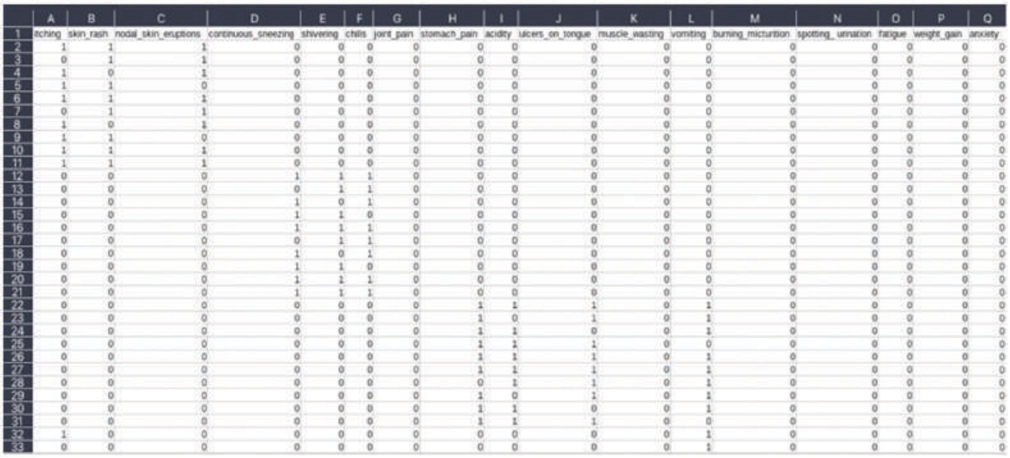
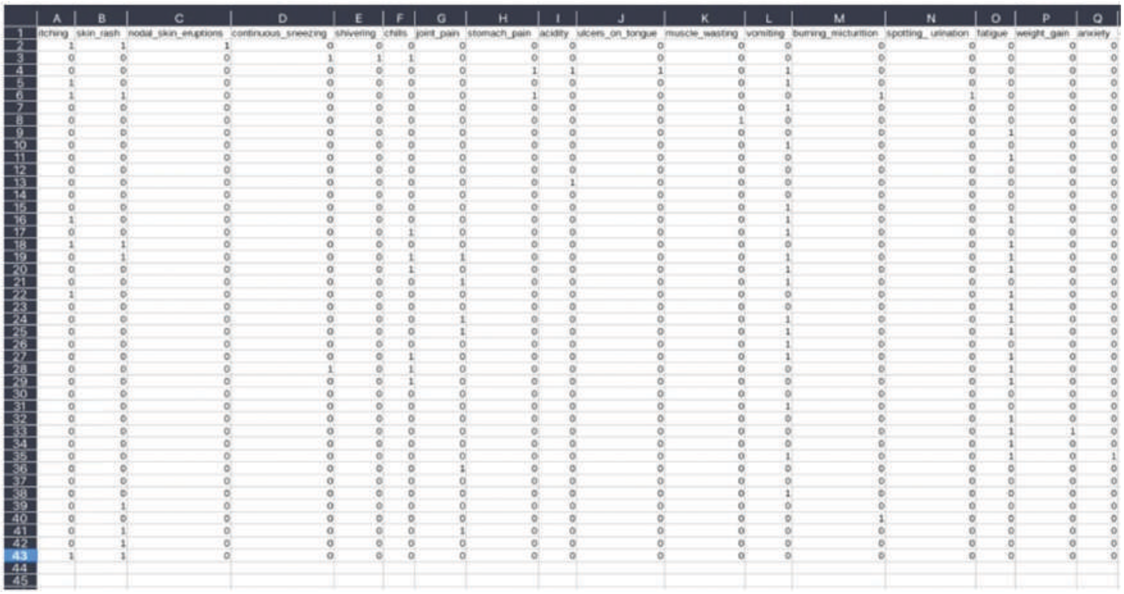


FIGURE 1 | Work flow.



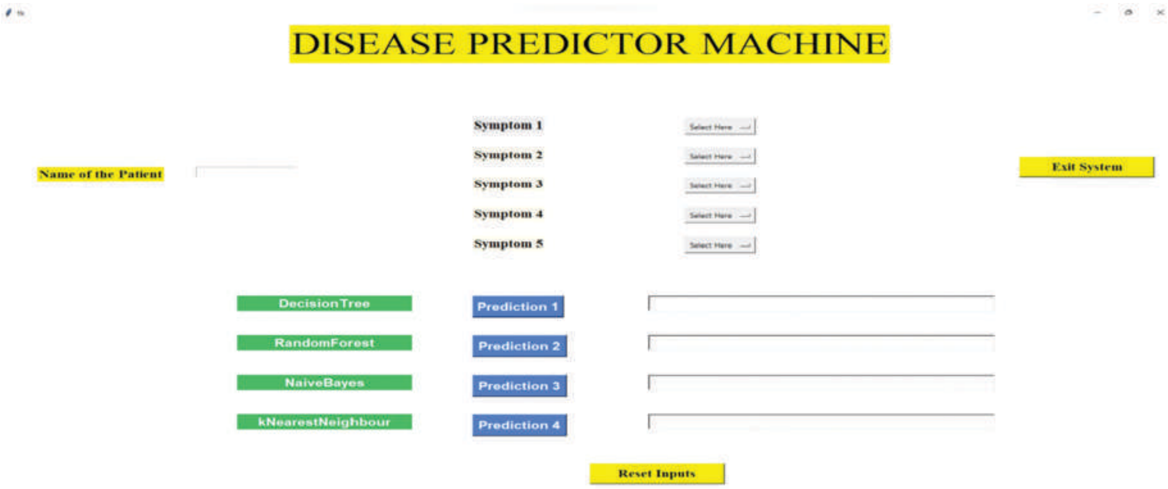
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_wasting	vomiting	burning_micturition	spotting_ urination	fatigue	weight_gain	anxiety
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
23	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
24	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
25	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
27	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
28	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
29	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
30	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
31	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
32	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

FIGURE 2 | Training set.



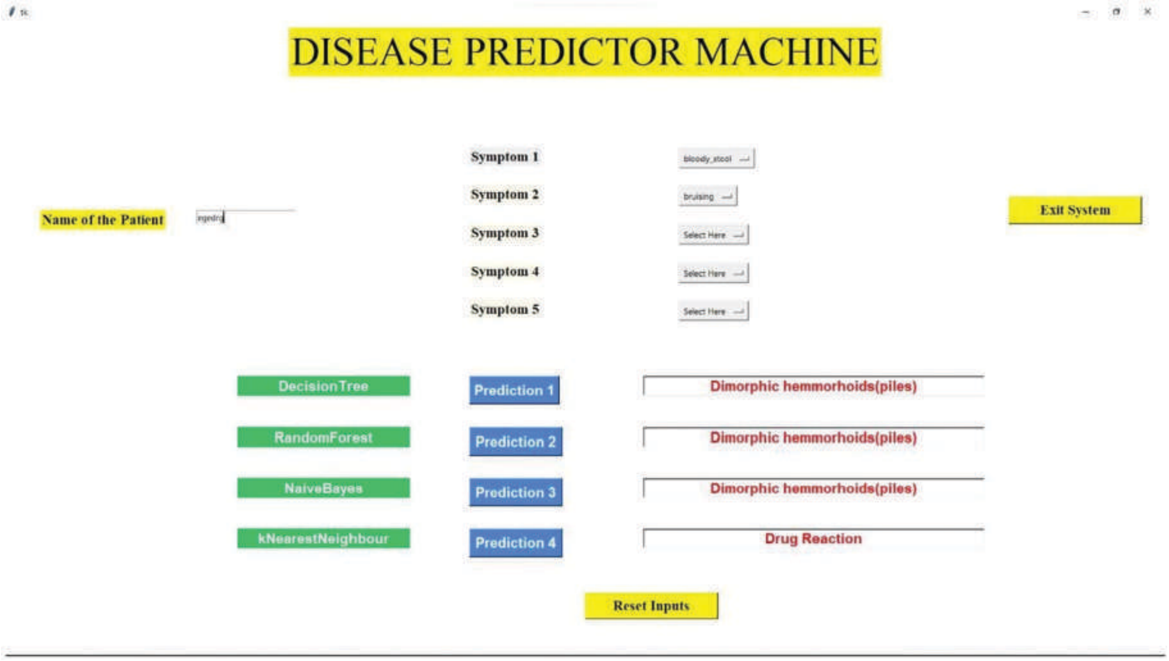
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_wasting	vomiting	burning_micturition	spotting_ urination	fatigue	weight_gain	anxiety
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
6	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
17	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
19	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0
20	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0
21	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
24	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
25	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
27	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0
28	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
29	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
35	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
36	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
39	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
41	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
42	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44																	
45																	

FIGURE 3 | Test set.



Symptom 1 Select Here
Symptom 2 Select Here
Symptom 3 Select Here
Symptom 4 Select Here
Symptom 5 Select Here

FIGURE 4 | Interface for prediction of disease.



Name of the Patient irgedrd

Symptom 1 bloody_stool
Symptom 2 bruising
Symptom 3 Select Here
Symptom 4 Select Here
Symptom 5 Select Here

Prediction 1 Dimorphic hemmorhoids(piles)
Prediction 2 Dimorphic hemmorhoids(piles)
Prediction 3 Dimorphic hemmorhoids(piles)
Prediction 4 Drug Reaction

FIGURE 5 | Final result.

(1) We need to load the required data.
(2) We need to calculate the distance between points, which is called the Euclidean distance.
(3) We have top k top distances.

Python was chosen for a variety of reasons. It is dependent on your perspective and background. It is made for programmers. One of the most well-known programming languages is Python. Python is one of the easiest programming languages to learn. It is quite simple, and we can use the grammar language in it as syntax. Python is one of the high-level languages, which has an inbuilt garbage collector that is used to free up the memory from the elements that are not used in the code.

Bits and pieces together

This approach can utilize the already done work by utilizing it as a starting point. All the information from accomplished work can be combined.

Equations

The equations should be inserted in an editable format from the equation editor.

$$f(x) = ao + \sum_{n=1}^{\infty}\left(b_n cos\frac{n\pi x}{s} + c_n sin\frac{n\pi x}{s}\right)$$

Proposed system

In this model, we (GUI) take the symptoms from the user and predict the disease he is suffering from. The interface responds immediately in a fraction of a time with accurate accuracy.

- The user has to fill in the details like his name.
- The user has to enter the symptoms of suffering, at least 2 symptoms.
- The system will store the data like his name and the disease he is suffering from so that treating him the next time will be easy and fast to cure him.

Methodology

A methodology is a representation of a system's structure, behavior, and other features. A system architecture is made up of system components and subsystems that interact to form the total. Individuals use an architecture diagram to abstract the overall structure of a software system and define constraints, linkages, and boundaries between components. The methodology of the work is shown in **Figure 1**.

Python has many applications. Some of them are the following:

- Web Development
  Many web development projects use Python because Python has introduced a lot of frameworks that make work easier, simpler, and more attractive.
- Data Science
  Data science itself involves so many stages like data mining, data sorting, data processing, etc. So, Python provides inbuilt functions that make work easier and simpler to work with.

Results

This dataset was acquired from a Kaggle reference. Here, in the dataset, we have 5,000 rows of data that help in training models very efficiently shown in **Figure 2**.

The testing data has nearly 45 rows that help in calculating accuracy shown in **Figure 3**.

Figure 4 shows the interface of the disease prediction scenario, and **Figure 5** shows the final result achieved after providing symptoms in the interface.

The work is being done with four machine learning algorithms, i.e., decision trees, random forest, KNN, and Naive Bayes. The best result was achieved with a random forest algorithm. The comparison of all classifiers is shown in **Figure 6**.

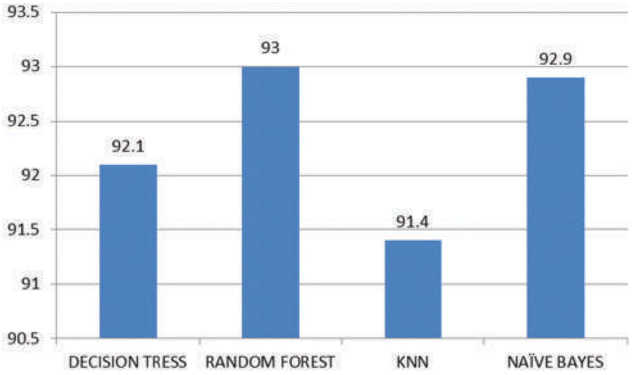


FIGURE 6 | Comparison with different machine learning models.

Conclusion

After completing the work, we can conclude that the random forest predicts the disease with high accuracy, and after the random forest, it is the decision tree that gives one of the best accuracies. We have created a system that can decrease the rush at hospitals and medical areas, and it also helps in reducing the workload on the medical staff. As a result, our system benefits both patients and the medical field. By building such types of systems, we can save time and money spent by the patients to undergo tests or scanning to know what they are suffering from. On average, our system achieved an accuracy of 93% in editing diseases with the symptoms given by the user with the random forest algorithm. In creating this system, we also added a way to store the data entered by the user in the database, which can be used in the future to help in creating a better version of such a system.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

References

1. Mohapatra H. HCR (English) using neural network. *Int J Adv Res Innov Ideas Educ.* (2015) 1:379385.
2. Mohapatra H, Rath AK. Detection and avoidance of water loss through municipality taps in India by using smart taps and ICT. *IET Wireless Sens Syst.* (2019) 9:447–57.
3. Mohapatra H, Rath AK. Fault tolerance in WSN through PE-LEACH protocol. *IET Wireless Sens Syst.* (2019) 9:358–65.
4. Mohapatra H, Debnath S, Rath AK. Energy management in wireless sensor network through EB-LEACH (No. 1192). *Easy Chair.* (2019).
5. Nirgude V, Mahapatra H, Shivarkar S. Face recognition system using principal component analysis & linear discriminant analysis method simultaneously with 3d morphable model and neural network BPNN method. *Glob J Adv Eng Technol Sci.* (2017) 4:1–6.