The 1st International Conference on Innovations in Computer Science & Software Engineering, (ICONICS 2016), December 15-16, 2016

Analyzing Diabetes datasets Using Data Mining Tools

Anita Shaikh
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
aneeta_2013@live.com

Sohail Abdul Sattar
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
sattar@neduet.edu.pk

Warda Tariq
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
warda_tariq@rocketmail.com

Anaum Hamid
Department of Computer Science and Information Technology,
NED University,
Karachi, Pakistan.
anaumhamid@gmail.com

***Abstract*— Analyze, examine, explore and to make use of data this we termed as data mining .Data mining is useful in various fields for eg in medicine and we may take help for predicting the non-communicable diseases like diabetics. Diabetes mellitus placed 4th among NCDs, caused 1.5 million global deaths each year worldwide [1]. We are using different classifying algorithms such as Naïve bayes , MLP, J.48, ZeroR, Random Forest, Regression to depict the result and compare them and our aim is to find solution to diagnose the disease by getting meaningful result out of the data**

Keywords—Data mining; Classification; Algorithm;Diabtes MelitusType II.

I. INTRODUCTION

Non-communicable diseases (NCDs) which include stroke, heart disease, cancer, chronic lung cancer and diabetes they together are responsible for almost 70% of the deaths worldwide [1] in which Diabetes mellitus Type ii is most common in all.

The number of patients suffered has quadrupled since 1980.it is estimated that 422 million people have diabetes all over the world and this figure may get doubles in the next 20 years. [1]

The top 10 countries which are affected are India, China, USA, Indonesia, Japan, Pakistan, Russia, Brazil, Italy and Bangladesh.[3]

About seven million Pakistanis had diagnosed Type ii diabetes mellitus it is estimated that in 2035, the figure will reach up to 12 million. [4]

In this situation, we need to look into the facts and figure and the risk factors involved in it.

This paper meant to be written to give us an idea so we can early on diagnose the disease by comparing different data mining algorithm.

II. BACKGROUND

Data mining is a process in which we are allowed to ascertain the patterns in the provided datasets by simply applying combination of methods like artificial intelligence, machine learning, statistics and database system.

Our objective is to obtain information from the dataset and alter it to a more meaningful structure that is understandable to a layman.

The data mining tool that we are opting for is WEKA. It gives us the ideas as it contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also suited for developing new machine learning schemes.[2]

In this particular example, we are using different classifiers which include naïve bayes, decision tree and regression techniques and neural networks to get the best results out of it.

III. METHODS

A. Classification

Naiive Bayes
ZeroR
MLPUnits

B. Regression

Logistic Regression

C. Decision Tree

J48
Random Forest

IV. DATASETS

The datasets had been taken from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases these datasets includes records of 768 patients, out of which 500 tested negative while 268 of them were tested positive.[9]

In this, we are analyzing the data with these nine Attributes that helps us to understand the possible prediction of this disease and which of the algorithm is more suitable for it.

The first eight attribute are the inputs set as input and the ninth attribute is the result which is used as a target which either “Positive” or “Negative”.

TABLE I. DATASETS OF DIABETIC PATIENTS

S.NO	Name	Description	Unit	Value range
01	Preg	No of Times Pregnant	Numeric value	0-9
02	Plas	Plasma GulocoseConcentrartion	Numeric value	0-199
03	Press	Diastolic Blood Pressure	mmHg	0-122
04	Skin	Triceps skin folds thickness	mm	0-99
05	Insulin	2-Hours Serum Insulin	mu/Uml	0-846
06	Mass	Body Mass Index	Weight in kg Height in m^2	0-67.1
07	Pedi	Diabets Pedigree Function	Numeric value	0.08-2.42
08	Age	Age	Numeric value	21-81
09	Classs	Diabetes Melitis Type II	Numeric value	Postive =1 ,Negative = 0

TABLE II. TABLULAR VIEW OF DATASETS

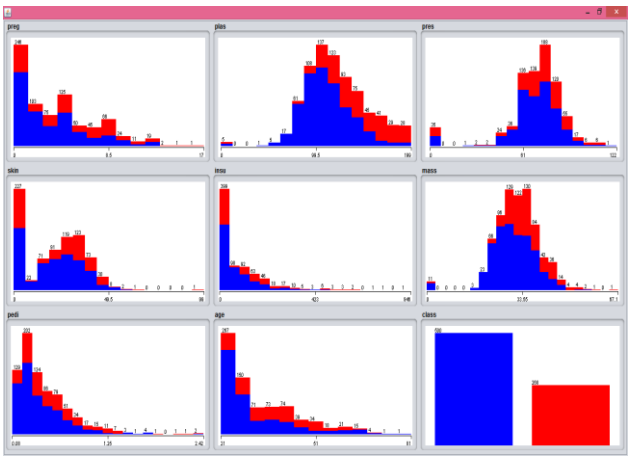
	preg	plas	pres	skin	insu	mass	pedi	age	class
1	6	148	72	35	0	33.6	0.627	50	tested_positive
2	1	85	66	29	0	26.6	0.351	31	tested_negative
3	8	183	64	0	0	23.3	0.672	32	tested_positive
4	1	89	66	23	94	28.1	0.167	21	tested_negative
5	0	137	40	35	168	43.1	2.288	33	tested_positive
6	5	116	74	0	0	25.6	0.201	30	tested_negative
7	3	78	50	32	88	31	0.248	26	tested_positive
8	10	115	0	0	0	35.3	0.134	29	tested_negative
9	2	197	70	45	543	30.5	0.158	53	tested_positive
10	8	125	96	0	0	0	0.232	54	tested_positive
11	4	110	92	0	0	37.6	0.191	30	tested_negative
12	10	168	74	0	0	38	0.537	34	tested_positive
13	10	139	80	0	0	27.1	1.441	57	tested_negative
14	1	189	60	23	846	30.1	0.398	59	tested_positive
15	5	166	72	19	175	25.8	0.587	51	tested_positive
16	7	100	0	0	0	30	0.484	32	tested_positive
17	0	118	84	47	230	45.8	0.551	31	tested_positive
18	7	107	74	0	0	29.6	0.254	31	tested_positive
19	1	103	30	38	83	43.3	0.183	33	tested_negative
20	1	115	70	30	96	34.6	0.529	32	tested_positive
21	3	126	88	41	235	39.3	0.704	27	tested_negative
22	8	99	84	0	0	35.4	0.388	50	tested_negative
23	7	196	90	0	0	39.8	0.451	41	tested_positive
24	9	119	80	35	0	29	0.263	29	tested_positive
25	11	143	94	33	146	36.6	0.254	51	tested_positive
26	10	125	70	26	115	31.1	0.205	41	tested_positive
27	7	147	76	0	0	39.4	0.257	43	tested_positive
28	1	97	66	15	140	23.2	0.487	22	tested_negative
29	13	145	82	19	110	22.2	0.245	57	tested_negative
30	5	117	92	0	0	34.1	0.337	38	tested_negative
31	5	109	75	26	0	36	0.546	60	tested_negative
32	3	158	76	36	245	31.6	0.851	28	tested_positive

V. GRAPHICAL REPRESENTATION OF ATTRIBUTES

Test Positive= Red

Test Negative =Blue

Fig. 1. Decision Tree (Random Forest)



Classification Algorithm and Their Evaluation

Output Prediction

These results were based on 90% percentage split In this table the column "Actual " and "predicted" represents the original result versus the predicted result however the column "error" represents the prediction error.

TABLE III. COMPARISON OF PREDICTION OF FIRST TWO INSTANCES BY USING DIFFERENT ALGORITHMS

N= tested_negative
P= tested_positive

Decision Attributes	Logistic Regression	Naïve Bayes	ZeroR	J.48	MLP	Random Foest
Instance 1						
Actual	N	N	N	N	N	N
Predicted	N	N	N	N	N	N
Prediction (True/False)	True	True	True	True	True	True
Instance 2						
Actual	P	P	P	P	P	P
Predicted	P	N	N	P	N	N
Prediction (True/False)	True	False	False	True	False	False

A. Naïve Bayes

This algorithm is named after Thomas Bayes who proved the bayes theorem. Naive Bayes is suitable in our situation in which we need to identify the possibilities of how many people are more prone towards diabetes.
This algorithm works on probability distribution function.

TABLE IV. PREDICTION TABLE

inst#	actual	predicted	error
1	1:tested_negative	1:tested_negative	0.99
2	2:tested_positive	1:tested_negative	+0.67
3	1:tested_negative	1:tested_negative	0.501
4	1:tested_negative	1:tested_negative	0.825

In Error column 0.99 means there is 99% chance of that instance to test negative which is true and 1% possibility that the instances could test positive.

"+ "means prediction came out untrue. However, in the second instance 67% chance for the instance to test negative as compared to the instance in which we have 99% surety hence it proved wrong.

0.67 is not to close to 0.99 which gives the algorithm a benefit of doubt as to predict positive or negative.
Zero R

ZeroR is the simplest classification method. It is that type of classification method which would lean on the target and ignore other attributes.

TABLE V. PREDICTION TABLE

inst#	actual	predicted	Error prediction
1	1:tested_negative	1:tested_negative	0.648
2	2:tested_positive	1:tested_negative	+0.648
3	1:tested_negative	1:tested_negative	0.648
4	1:tested_negative	1:tested_negative	0.648
5	2:tested_positive	1:tested_negative	+0.648

We always get the same result for every instance either 65% (0.352 test negative) or 35% (0.352 test positive)

B. Logistic Regression

Logistic regression was developed by statistician David Cox in 1958. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. [7]

TABLE VI. PREDICTION TABLE

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.981
2	1:tested_ positive	1:tested_ positive	0.517
3	1:tested_negative	1:tested_ positive	+0.5
4	1:tested_negative	1:tested_negative	0.721
5	1:tested_ positive	2:tested_positive	0.582
6	1:tested_negative	1:tested_negative	0.841
7	1:tested_ positive	2:tested_positive	0.921
8	1:tested_negative	2:tested_ negative	0.927

C. Random Forest

Random forest generates many single classification trees. To classify a new object from an input, put the input vector down each of the trees in the forest. Each tree generates their own results and then they select one set of a class. [15]

```
plas< 111.5
|  preg< 7.5
|  |  skin < 29.5
|  |  |  age < 30.5
|  |  |  |  skin <19.5 :tested_negative (122/0)
|  |  |  |  skin >= 19.5
|  |  |  |  |  plas<94.5 :tested_negative (40/0)
|  |  |  |  |  plas>= 94.5
|  |  |  |  |  |  mass <32.7 :tested_negative (18/0)
|  |  |  |  |  |  mass >= 32.7
|  |  |  |  |  |  |  preg<0.5 :tested_positive (2/0)
|  |  |  |  |  |  |  preg>= 0.5
|  |  |  |  |  |  |  |  skin <23.5 :tested_positive (1/0)
|  |  |  |  |  |  |  |  skin >= 23.5 :tested_negative (10/0)
|  |  |  |  |  pedi>= 0.22
|  |  |  |  |  |  mass <37 :tested_positive (15/0)
|  |  |  |  |  |  mass >= 37
|  |  |  |  |  |  |  pres< 89
|  |  |  |  |  |  |  |  skin <36.5 :tested_negative (5/0)
|  |  |  |  |  |  |  |  skin >= 36.5 :tested_positive (2/0)
|  |  |  |  |  |  |  pres>= 89 :tested_positive (3/0)
|  |  |  plas>= 146.5 :tested_positive (48/0)

Size of the tree : 189
```

TABLE VII. PREDICTION TABLE

inst#	actual	predicted	Error prediction
1	1:tested_negative	1:tested_negative	0.58
2	1:tested_negative	1:tested_negative	0.55
3	1:tested_negative	1:tested_negative	0.95
4	1:tested_negative	1:tested_negative	1
5	1:tested_negative	1:tested_negative	0.6
6	1:tested_negative	1:tested_negative	0.81
7	1:tested_negative	2:tested_positive	+0.83
8	1:tested_negative	2:tested_positive	+0.65

D. Multilayer Perception

It works on how different attributes results process and interact with one another and alter their results in such a way that the final outcome is the filtered through each node (neuron).

Multi-Layer perception bestows great advantages as it is used for pattern classification, recognition, prediction and approximation.

TABLE VIII. PREDICTION TABLE

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.692
2	1:tested_negative	1:tested_negative	0.583
3	1:tested_negative	1:tested_negative	0.942
4	1:tested_negative	1:tested_negative	0.954
5	1:tested_negative	1:tested_negative	0.945
6	1:tested_negative	1:tested_negative	0.894
7	1:tested_negative	2:tested_positive	+80.85
8	1:tested_negative	2:tested_positive	+0.55

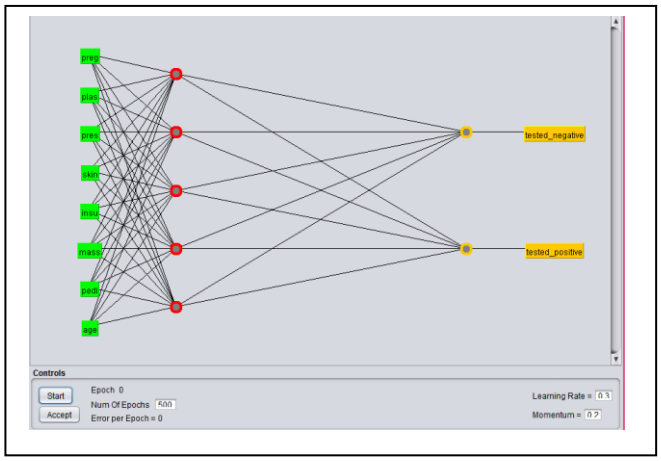


Fig. 2. Neural network (MLP)

Here we can see a network of different layers namely input layer, hidden layer and output layer consisting of input nodes

(green) or "neurons", output nodes (yellow) and some hidden nodes(red) some of them are visible. The nodes in the network are all sigmoid .Each connected network has some value in it which will be pass on to other nodes and each nodes perform a weighted sum of its input and pass it on until we get some results.Hidden layer depends upon the complexity of the data.[5].

Here MLP does show result with minimum error rate but it processes slow as compared to others.

E. J.48

Jr8 is basically an implementation of C4.5 algorithm [6] J48 decision tree decides which attributes is the most decisive one and which one is least and over and then these attributes further divided into sub tree. It generates a binary tree, unlike Random Forest decision tree. It use the concept of entropy, difference in entropy gives us the attribute which is free to make decisions.

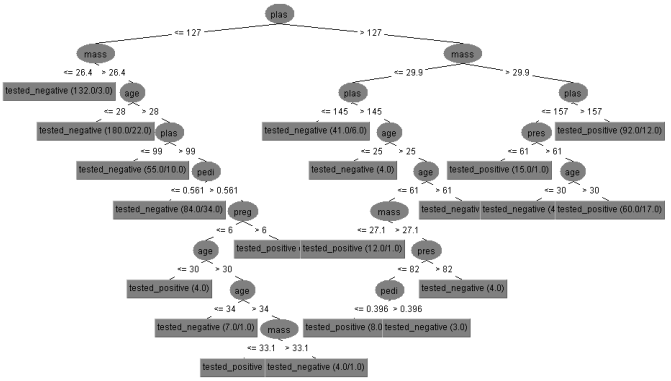


Fig. 3. Descion tree (j48)

TABLE IX. PREDICTION TABLE

inst#	actual	predicted	error prediction
1	1:tested_negative	1:tested_negative	0.982
2	2:tested_positive	2:tested_positive	0.635
3	1:tested_negative	2:tested_positive	+0.635
4	1:tested_negative	1:tested_negative	0.867
5	2:tested_positive	1:tested_negative	+0.9
6	1:tested_negative	1:tested_negative	0.867

VI. CLASSIFICATION RESULTS

Positive = identified

Negative = rejected.

Therefore:

TP=True positive = correctly identified

FP=False positive = incorrectly identified

TN=True negative = correctly rejected

FN=False negative = incorrectly rejected.[6]

Accuracy= (TP + TN) / (TP + FP + TN + FN) [8]

TABLE X. COMPARISON OF ACCURACY BETWEEN DIFFERENT ALGORITHMS

S.No	Classification Type	Accuracy (%)
1	Naïve Bayes	76.3 %
2	MLP	81.8182%
3	J.48	75.3%
4	ZeroR	67.5%
5.	Random Forest	79.2%
6.	Regression	76.8%
7.	Logistic Regression	79.2%

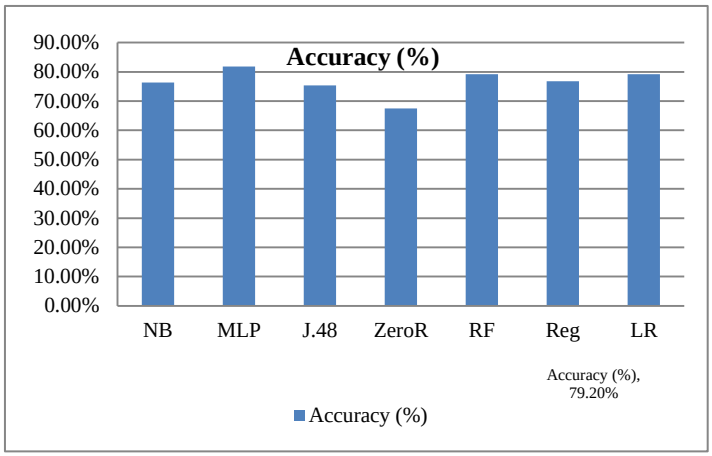


Fig. 4. Graphical representation of Accuracy over different algorithm.

VII. CONFUSION MATRIX

TABLE XI. COMPARISON OF CONFUSION MATRIX BETWEEN DIFFERENT ALGORITHMS

Algorithm	TN	FP	FN	TP
Naïve Bayes	45	7	10	15
MLP	49	3	11	14
J.48	43	9	10	15
ZeroR	52	0	25	0
Random Forest	45	7	9	16
Logistic Regression	45	7	9	16

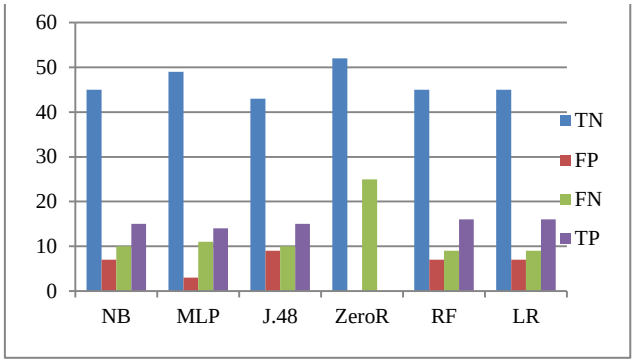


Fig. 5. Graphical representation of Confusion Matrix over different algorithm.

TABLE XII. COMPARISON OF ABSOLUTE ERROR BETWEEN DIFFERENT ALGORITHMS

S.No	Error
Naïve Bayes	0.2841
ZeroR	0.4481
J48	0.3036
Logistic	0.2867
Random Forest	0.3884
MLP	0.244

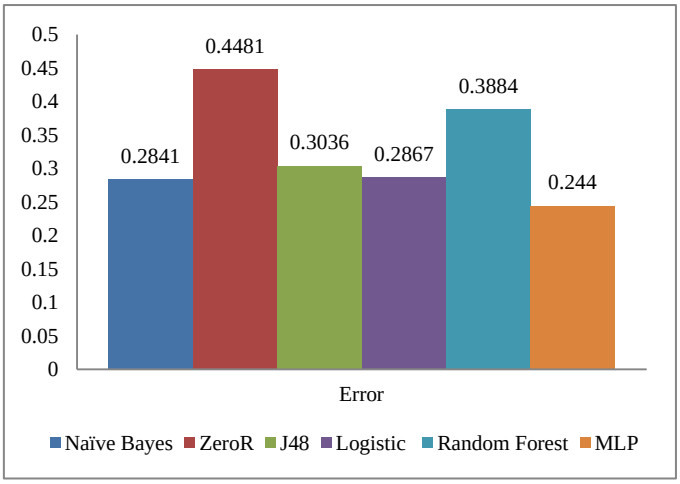


Fig. 6. Graphical representation of absolute error over different algorithm.

VIII. CONCLUSION AND FUTURE SCOPE

In order to make effective and efficient results, we need to work on a different algorithm and to make sure which suits best. Diagnosing diabetes through data mining tool over medical records of patients though it has been done by a majority of the researchers but the more we dig into the matter the furthest we can go.

In terms of performance we find multi layer perception function most effective hence it shows fewer errors however it takes too much processing time because it requires calculation of weights of each node. ZeroR is useful to determine baseline performance for others classification method. Naïve Bayes is also very efficient as it gives a predominant result after each validation but its performance is not quit impressive. J4.8 gives a graphical image of the precedence of the attribute as it calculates the priority of each attribute with other and yet it also predicts accurate results with least error hence it requires time.

Our objective of comparing the algorithm on the same dataset, analyzing and predicting the results out of it has been achieved.

In future, we will be interested in gathering information among our own neighborhood and we were keen to get new results which lead us toward more precise and accurate divination.

REFERENCES

[1] WorldHealth Organization, Diabetes Programm http://www.who.int/diabetes/en/

[2] Machine Learning Group at the University of Waikato..Weka 3: Data Mining Software in Java. Retrieved September 4, 2016 from http://www.cs.waikato.ac.nz/ml/weka/

[3] Sanofi, Diabetes Pakistan, Statistics http://www.sanofidiabetes.com.pk/web/about_diabetes/statistics

[4] The News International https://www.thenews.com.pk/print/73051-seven-million-pakistanis-suffering-from-type-2-diabetes

[5] Ian H. Witten, Department of Computer Science University of Waikato New Zealand, " Simple neural networks", "More Data Mining with Weka".More Data Mining with Weka, Simple Neuarl Network, https://drive.google.com/file/d/0B-f7ZbfsS9-xcEFUZ095UUpnVlU/edit

[6] Sathees Kumar B , Gayathri P,Department of Computer Science,Bishop Heber College ,"Analysis ofAdult-Onset Diabetes Using Data Mining Classification Algorithms", International Journal of Modern Computer Science(IJMCS)ISSN: 2320-7868 (Online)Volume No.-2, Issue No.-3, June, 2014Conference proceeding.

[7] Logistic Regression https://en.wikipedia.org/wiki/Logistic_regression

[8] P. Radha , Dr. B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.

[9] PimaIndiansDiabetesDataSethttps://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[10] Aiswarya Iyer, S. Jeyalathaand Ronak Sumbaly "Diagnosisof Diabetes Using Classification Mining Techniques" , International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[11] N.Satyanandam,Dr.Ch. Satyanarayana,Md.Riyazuddin, Amjan.Shaik "Data Mining Machine Learning Approaches and Medical Diagnose Systems ", International Journal of Computer & Organization Trends – Volume2Issue3-2012

[12] Sadri Sa'di, Amanj Maleki, Ramin Hashemi, Zahra Panbechi and Kamal Chalabi, "Comparison Of Data Mining Algorithms In The Diagnosis Of Type II diabetes", International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015

[13] D. Ezaz Ahmed, Dr. Y.K. Mathur, Dr Varun Kumar, "Knowledge Discovery in Health Care Datasets Using Data MiningTools", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 3, No.4, 2012117.

[14] Sukhjinder Singh, Kamaljit Kaur, "A Review on Diagnosis of Diabetes in Data Mining", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438.

[15] Tahani Daghistani,Riyad Alshammari, "Diagnosis of Diabetes by Applying Data Mining Classification TechniquesComparison ofThree Data Mining Algorithms", IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 7, No. 7, 2016.