# Standard Cog Exploration on Medicinal Data

Vadamodula Prasad
Associate Professor & Head
Dept of CSE
Raghu Institute of Technology

Thamada Srinivasa Rao
Associate Professor
Dept of CSE
Gitam Institute of Technology

Ankit Kumar Surana
Systems Engineer
INFOSYS
Mysore Campus

## ABSTRACT
In this paper, we diagnose disease by implementing principal component analysis based on the user entered values for various symptoms. It considers components from the user entered values and compares them with the values in the database. Here the diagnosis is done by the method of prediction using trained data sets (TDS) and the results are compared by using suitable data matching systems (DMS).The TDS are provided by Intelligent System Laboratory of K.N. Toosi University of Technology, Imam Khomeini Hospital. If the user entered symptom values exist in the medical dataset then the percentage of getting an output that is true is 100%, if the symptom values do not exist, then the disease is predicted. PCA provides a precise disease prediction with the disease name and referral source which are optimized and neared to the true value in the datasets. Prediction based on machine learning algorithm gives an output that is 92%. This work predicts the actual levels of thyroid in human body.

## General Terms
Hyperthyroid disease ,Prediction , Machine learning algorithms.

## Keywords
Data matching system, Database, Prediction, Principal component analysis, Trained datasets.

## 1. INTRODUCTION
Hyperthyroid[1,7] may not cause identifiable symptoms. More often, however, the symptoms are thwarting, disabling or even critical. The contemporary medical diagnosis system is a system that needs a lot of technical encroachment mainly in advisory system. It is an offline browsing system which is being operated by human experts where there are loads of difficulties. This has been practiced for many years and in spite of pain the patient remains unfamiliar of the disease he is suffering from. The other drawback is that it can't be carried out in everywhere due to lack of professionals. There is also a chance that due to human experts [16] knowingly or unknowingly there may be some errors. Educated persons using this website will know the disease which they are suffering by entering the symptoms. The doctors can also use this website to access the information stored in web portal for upgrading their skills and also diagnose using TDS. The others who can use this system are Hospitals & other stake owners connected with Health Science will be benefited with this web portal to increase their efficiency and effectiveness. The information system developed in this project consists about modules like Human Health Concerns, Human Diseases & surveillance program.

The system[7][10] is maintained with both string harmonizing and optimization techniques in order to provide a closest disease, what happens is, if the user entered the disease symptoms in the input form given by the expert, it displays the actual disease that the patient is suffering with, if the string matches, or else it shows the content which is a value that is not the closest disease, hence giving deceptive disease to the user, for this purpose we implemented a optimization technique which will calculate the probability of occurrence of the closest disease.

## 2. EXISTING SYSTEM
In the existing system people have to visit the doctor even for small issues related to health. They are depended on doctors even after having their reports on various tests generated. Though there are expert systems available to resolve this issue there is some glitch in obtaining the report from those expert systems i.e., even after the implementation of an expert system for predicting diseases[15,16] , not all the diseases are predicted accurately. The user enters the symptoms that he is suffering from and hopes to get a result but due to insufficient data in the existing knowledge base there arises two cases where a report might be generated if there exists any report for the user entered symptoms else no report is generated. This has been further resolved by using some optimization techniques but none gives an accurate and a closest optimized result. Thus making the scope of the system to be used only by those whose data is actually present in the knowledge base if not giving some predicted result that is not related to the disease the patient is suffering from. Hence the main notion of predicting an accurate result for every entry by the user remains uncertain[11,12,13]

## 3. PROPOSED SYSTEM
Health is not valued till sickness comes and with technology[6] becoming more and more advanced and people getting more and more acquainted with web and the resources available on it they have become more cautious regarding their health. The web gateways have thus become essential to provide health related solutions at ease to the user. This offline web gateway is an extension of the system that predicted the result that is not to close enough to the user entered values[8,9,10]. This offline web gateway is expected to have the following feature:

This web gateway consists of two sections named (i) Disease System[17], (ii) Predict Disease[18]. In these mentioned fields Disease System contains the static information about the diseases their causes and their symptoms and Predict Disease is an advisory part of the system which is a part of the web gateway that provides the user to get a health advice against the disease for the symptoms entered. It also provides a facility to contact the experts i.e., doctors regarding any

further issues in their health. Provides information about various diseases, their causes and their symptoms provided by the users. It not only removes the issue of patient attending the doctor but also resolves the issue of not getting the accurate results to user entered symptoms.

## 3.1 PCA

Principal Component Analysis[1][2][3][4][5] is a way of identifying patterns in given data and to precise the data in such a way that it highlight their likenesses and dissimilarities. Since determining patterns in data can be a tedious task and also in some cases very hard in high dimensional data, where the luxury of graphical representation is not available, PCA is powerful for analyzing data. The other main advantage of PCA is that once these patterns in data are found, you can compress the data, i.e. by reducing the number of dimensions, without much loss of information.

## 3.2 Implementation

The optimization model employed here implements PCA[19,20] that makes use of the set of observations for different components to predict the disease. The knowledge base might not contain all the user entered component values but for an expert system to be efficient and useful one an output must be produced. The produced output must be relevant to the user entered values i.e., it must be close to it. This model considers these set of observations for different components to predict the output closest to it in the knowledge base. If the value is present in the knowledge base then the relevant output is displayed. If the knowledge base does not contain the user entered components then we make use of this algorithm implemented.

This algorithm uses the input components given for calculating a value that can be used on knowledge base to predict the closest output. This is done by calculating sum of squares of these user entered component values and comparing it with the knowledge base values. The value that is closest to our comparison is the closest output. The main steps in the algorithm:

We consider the user entered symptom values as variable x and the result outcome from the sum of squares of these inputs as X.

This result X will be now compared to the value stored in database is indicated with a value Y.

Hence the tuple in the database with the least or minimum comparison value is displayed as the closest result or predicted disease for the user entered symptoms.

Let us consider different inputs from user as

$$x_1, x_2, x_3, x_4\ldots\ldots, x_n$$

where 'n' is the number of input fields to be entered by the user. We now calculate 'X' the sum of squares for all these inputs by multiplying it with a constant '€'.

$$X = \sum(x_1^a + x_2^a + x_3^a + x_4^a + \ldots\ldots\ldots + x_i^a) * €,$$

where $x_i$ are the symptoms values entered by the user and i=1 to n, n being the number of input symptoms to be entered and € is a constant and 'a' being the power value which can be any value ranging from a=2 to n, where n is always a positive integer value less than 100, Hence the above equation can also be written as X=TE,

where $T=\sum(x_1^a + x_2^a + x_3^a + x_4^a + \ldots\ldots\ldots + x_i^a) \& E=€.$

Similarly we consider output Y i.e. equal to the sum of powers of 'a', where a=2 to n, of the symptoms values multiplied with € in the knowledge base. The values X and Y are now compared and the least absolute value gives the closest and the relevant output value.

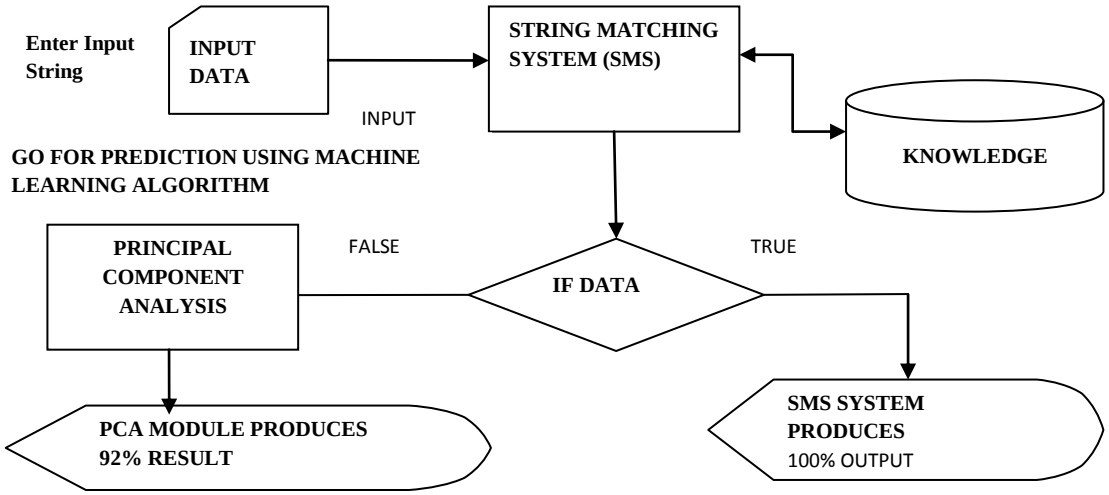
Predicted Result=Minimum{Y-X}.

# 4. SYSTEM DESIGN

In the development of this offline web portal we have implemented the below said

1) Data matching system using TDS

2) Event flow diagram



## 4.1 Event Flow Diagram

The process is started by manually entering the symptoms. These entered values are used to create a string that is used on the database for comparison. These string values are then crosschecked with the database string values if any string in the database matches with those entered string values then an accurate disease and the referral source is given as output to the user indicating the accurate output. If the values in the database do not match with the values entered then the produced algorithm is used to predict the disease and the referral source indicates adjusted output.

## 4.2 Knowledge Base

Knowledge Base is obtained from the Intelligent System Laboratory of K.N.Toosi University of Technology, Imam Khomeini hospital the following are the attributes taken into consideration for the TDS[20].

S1=Age S2=Gender
S3=On Thyroxin
S4=QueryOn Thyroxin
S5=Anti-Thyroid Medication S6=Sick
S7=Pregnant
S8=Thyroid Surgery
S9=I131 S10=Hypothyroid
S11= Hyperthyroid S12=Lithium
S13=Goiter
S14=Tumor
S15=Hypo pituitary S16=Psych
S17=TSH
S18=TSH Value
S19=T3 S20=T3 Value
S21=TT4 S22=TT4 Value
S23=T4U S24=T4U Value
S25=FTI S26=FTI Value
S27=TBG S28=TBG Value
Indication in Data Sets 1- Yes 0-No

## 5. RESULTS

***Continued to Figures 2(a) & 2(b)***

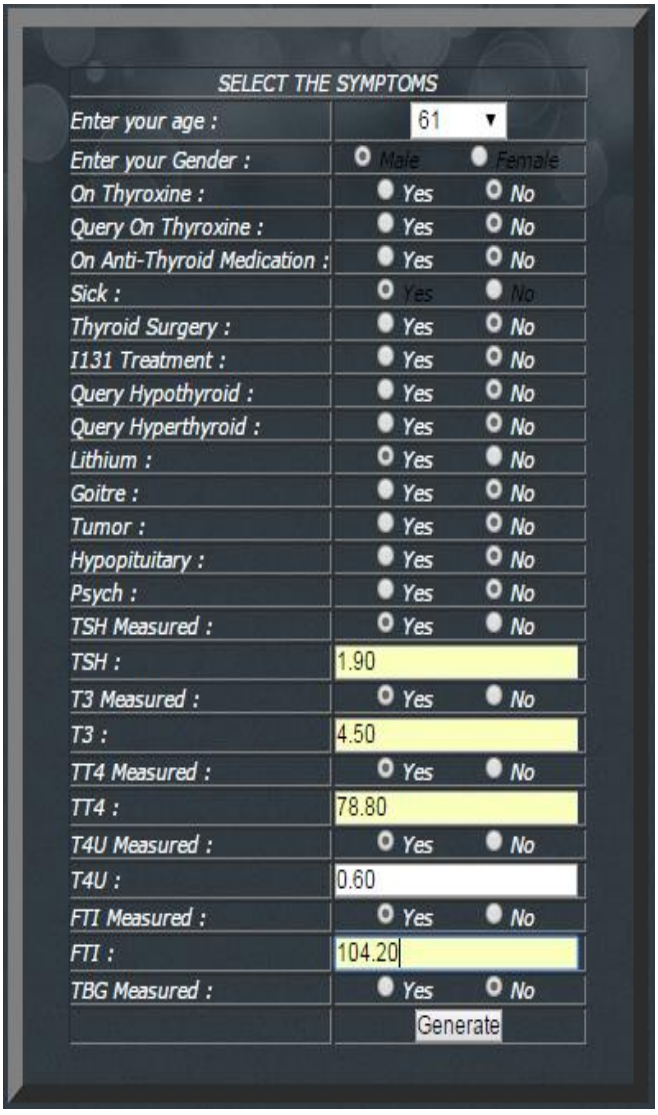



**Figure 2(a): Input screen to enter the symptoms and to identify the optimized disease based on the symptoms entered.**

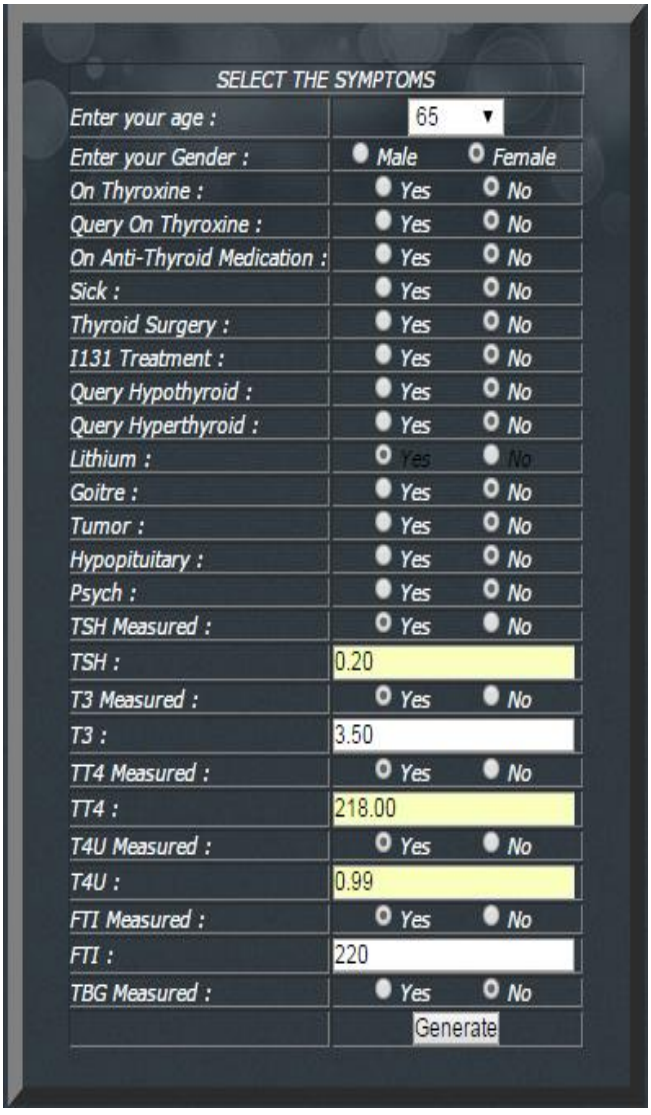



**Figure 2(b): Input screen to enter the symptoms and to identify the optimized disease based on the symptoms entered.**

Output for figure 2(a) obtained by implementing PCA:

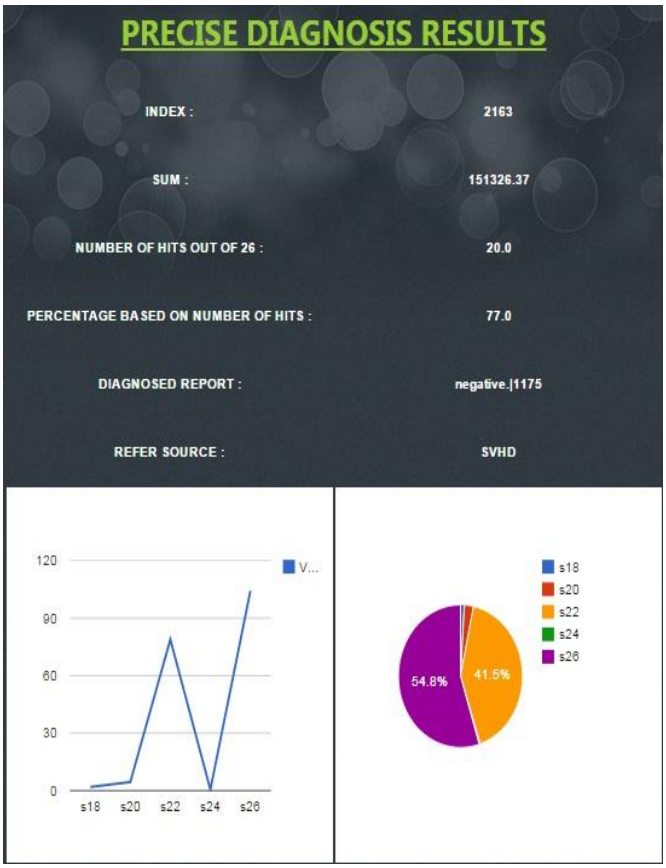
Disease:Negative.|1175

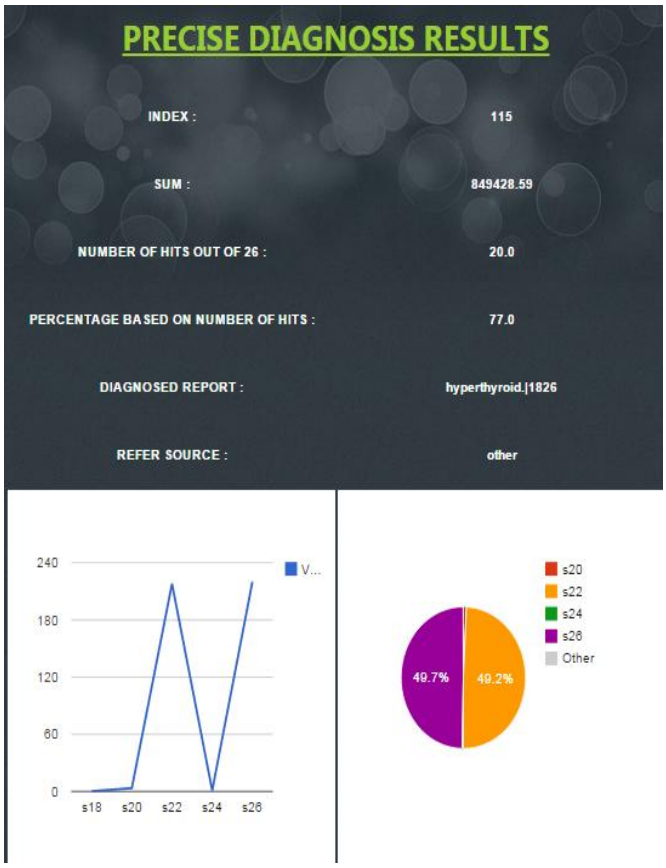
Referral Source:SVHD

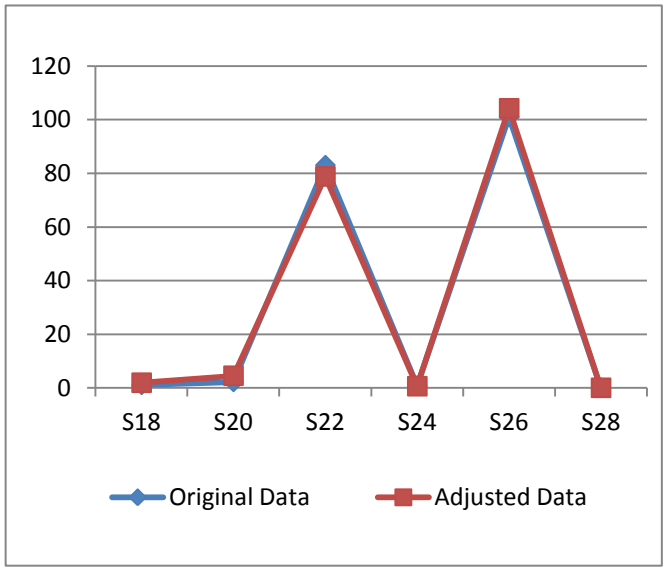
Output for figure 2(b) obtained by implementing PCA:

Disease: Hyperthyroid.|1826

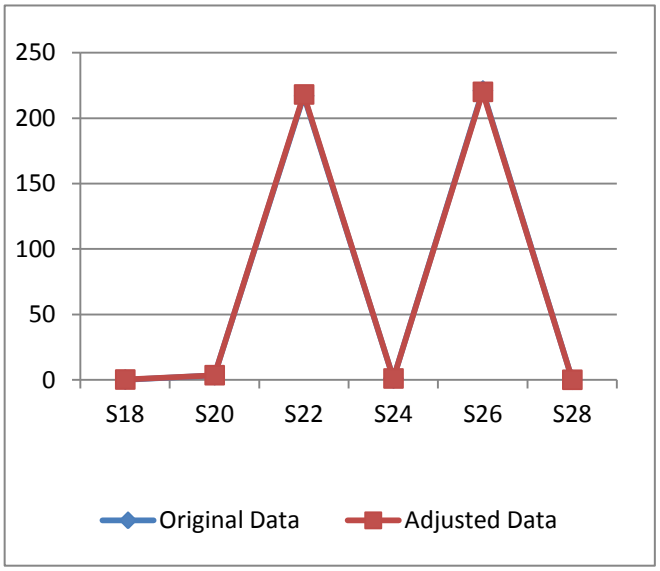
Referral Source: Other




***Graph**: The above graph represents the difference between the original data that has been predicted as the closest value in the database and the user entered values for figure 2(a) .*




***Graph**: The above graph represents the difference between the original data that has been predicted as the closest value in the database and the user entered values for figure 2(b).*

## 6. CONCLUSION

This project "Standard cog exploration on medicinal data " is an implementation of PCA where users can know about a disease , their causes and referral sources. By entering the symptom values, the users can also know about the disease they are suffering from. This system is developed using JSP and MYSQL database as backend.

Its main objective is to have a well-designed interface for giving health related knowledge and suggestions in the area of any disease field by providing the symptoms as an input to the system and interact with the data matching system and the user without the need of an expert at all times. This system predicts results with an accuracy ranging above 92% that is better than other techniques like ridge regression and lasso. We can also extend, design and develop the string matching systems for drug therapy for finding the right drug to cure the disease. By the thorough interaction with the users and beneficiaries, the functionality of the system can be extended further to many more areas in and around the world.

## 7. FUTURE ENHANCEMENTS

This approach is to find the disease using a machine learning which produces an output ,that has better percentages for different inputs compared to those produced by other algorithms like ridge regression, lasso. This system can be further extended to predict outputs for various other diseases by increasing the number of factors that are considered for predicting the output.

## 8. REFERENCES

[1] Poncelet, Pascal, Maguelonne Teisseire, Florent Masseglia: "Data Mining Patterns: New Methods And Application", Information Science Reference, Hershey PA, pp. 120-121(2008)

[2] Jolliffe, I.T: "Principal Component Analysis", Springer Verlag New York Inc. New York, pp. 7-26 (2002)

[3] K. Y. Yeung*,W. L. Ruzzo,"Principal component analysis for clustering gene expression data", Volume 17, Issue 19, Pages :763-774, doi:10.1093/bioinformatics/17.9.763

[4] Anton Buhagiar, Exploration and reduction of data using principal component analysis, A journal of University of Malta Medical School , Volume 14 , Issue 1 ,Page : 27 , 2002

[5] Dr. D.C. Barber[a], P.J. Howlett[a] & R.C Smart[a],Principal component analysis in medical research , DOI:10.1080/768371123, pages 39-43

[6] Prasad, V. "Plug in generator to produce variant outputs for unique data." *International Journal of Research in Engineering and Science (IJRES)* 2.4 (2014): 14-20.

[7] Prasad, V., T. SrinivasaRao, and M. Surendra Prasad Babu. "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms." *Soft Computing* (2015): 1-11.

[8] V Prasad, T. SrinivasaRao. "Health Diagnosis Expert Advisory System on Trained Data Sets for Hyperthyroid." *International Journal of Computer Applications (IJCA)* 102.3 (2014): 40-48.

[9] V Prasad ,KSravani , MdParvez, N Sekhar, G. Sireesha. "Human Motion Detection Using Passive Infra Red Sensor." *International Journal of Research in Computer Applications & Information © IASTER 2014* 2.2 (2014): 28-32.

[10] Prasad, V., T. SrinivasRao, and M. Surendra Prasad Babu. "Offline Analysis & Optimistic Approach on Livestock Expert Advisory System." *Artificial Intelligent Systems and Machine Learning* 5.12 (2013): 488.

[11] V Prasad, Dr T SrinivasaRao, Prof M. S. P. Babu. "TDD via Hybrid Architecture composing Rough Data Sets and ML Algorithms." *Second International Conference on Emerging Research in " Computing , Information, Communication and Applications" (ERCICA-2014),*. Vol. 1. No. 01. ELSEVIER, 2014.

[12] Prasad, V., Dr T. SrinivasaRao, and M. PurnachandraRao. "Proportional analysis of non linear trained datasets on identified test datasets." *International Conference on recent trends and research issues in computer science & engineering.*. Vol. 1. No. 1. ICRTRICSE-2k15, Dept of CSSE, Andhra University , Visakhapatnam, 2015.

[13] Prasad, Vadamodula, and TamadaSrinivasaRao. "Implementation of Regularization Method Ridge Regression on Specific Medical Datasets." *International Journal of Research in Computer Applications & Information Technology* 3.2 (2015): 25-33.

[14] Prasad, V., Dr T. SrinivasaRao, and B. Sai Ram. "Information Clustering Based Upon Rough Sets." *International Journal of Scientific Engineering and Technology Research (IJSETR)* 3.41 (2014): 8330-8333.

[15] K Vahini, V Prasad, U. V. Chandra Sekhar. "Defend Data using ELGAMAL Digital Signature Data Decryption Algorithm." *(IJCSIT) International Journal of Computer Science and Information Technologies* 5.4 (2014): 5062-5067.

[16] Expert System Based on Neural-Fuzzy Rules for Thyroid Diseases Diagnosis© Springer-Verlag Berlin Heidelberg 2012

[17] CROCHEMORE, CZUMAJ -Speeding up two string-matching algorithms -, et al. - 1994

[18] Thyroid Disease Diagnoses using C4.5 Algorithms and Data Mining Techniques ,2011

[19] Thyroid disease records supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia, 1987.

[20] Thyroid Disease Data Set. https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

[21] PCA-Principal Component Analysis. Dr. Saed Sayad, University of Toronto. 2010. http://chem-eng.utoronto.ca/~datamining/Presentations /PCA.pdf