Data analysis using Box and Whisker Plot for Lung Cancer

Chandrasegar Thirumalai, IEEE Member,
School of Information Technology and Engineering,
VIT University, Vellore, India.
chandru01@gmail.com

Vignesh M
MS Software Engineering,
School of Information Technology and Engineering
VIT University, Vellore, India.
vignesh2k18@gmail.com

Balaji R
MS Software Engineering,
School of Information Technology and Engineering,
VIT University, Vellore, India.
balaji.r2013@vit.ac.in

***Abstract*— In statistical analysis, we have a collection of data, with the use of these data, we have to do analysis based on our requirements. With the collection of data using Statistical analysis, we deal collection, analysis, presentation and organizing the data. With the help of statistical analysis, we can find underlying patterns, relationships, and trends between data samples. The R system for statistical computing is an environment for data analysis and graphics. Here we are going to implement boxplot method and control chart methods for Lung cancer dataset. With the help of boxplot, we can easily make relations between samples and we can find the outliers.**

Keywords-component; Data analysis, Lung Cancer, Decision making

I. INTRODUCTION

We have taken lung cancer datasets of 12 primary attributes as shown in the following Table I and II.

TABLE I. DATA SET OF 1ST PART OF LUNG CANCER ATTRIBUTES.

Age	Smoking status	Years smoked	Average per day	Gender	Grade
68	Smoker	10	15	Male	UG
77	Former Smoker	15	10	Male	PG
68	Non Smoker	0	0	Male	PG
71	Smoker	27	10	Male	Nil
74	Smoker	10	5	Male	Nil
51	Smoker	10	3	Female	UG
54	Former Smoker	14	6	Female	PG
50	Non Smoker	0	0	Female	Nil
60	Smoker	5	5	Male	UG
54	Smoker	12	5	Male	PG
54	Non Smoker	0	0	Male	UG
56	Former Smoker	12	12	Male	Nil
87	Smoker	10	10	Male	PG
45	Non Smoker	0	0	Male	PG
76	Former Smoker	25	12	Male	UG

To analyze the relevant data of Lung cancer dataset we have an applied Box plot data analysis method which is shown in Section 3. A boxplot is a data analysis method used to find the output of the samples. With the use of boxplot, we can easily compare the different datasets. In other words, boxplot also called box and whisker plot method.

TABLE II. DATA SET OF 2ND PART OF LUNG CANCER ATTRIBUTES.

Race	Height	Weight	Family history	Copd	Year	Cancer
Asian	175	85	No	Yes	2000	Yes
Asian	180	90	Yes	Yes	2001	Yes
Asian	182	57	Yes	No	2002	No
American Indian	170	80	Yes	Yes	2003	Yes
African American	182	85	No	Yes	2000	No
White	170	60	Yes	Yes	2002	No
Latin	175	65	No	No	2003	No
Asian	178	59	Yes	No	2004	No
American Indian	187	70	No	No	2005	No
American Indian	187	54	Yes	Yes	2002	No
American Indian	187	56	Yes	Yes	2003	No
Asian	187	58	Yes	Yes	2001	Yes
Asian	185	89	Yes	Yes	2003	Yes
Asian	185	84	No	Yes	2002	No
Asian	185	74	No	Yes	2004	Yes

This instructive datasets are used as the commitment to figure the Pearson [6], [11], [14], [16], [22], [24]. In the present days, there are enormous measures of data recorded by the banks and exploring them requires complex estimations. We played out the item metric examination on the given enlightening accumulation. From the data examination [8], [9], [12], [17], [18], [20] we can pick which quality can be viewed

978-1-5090-5682-8 /17/$31.00 ©2017 IEEE

as and which trademark can be expelled. For instance, in the Pearson strategy if the estimation of r is more than 0.5 then the credits are thought to be unequivocally related and if it is underneath 0.3 the qualities are pitifully related.

A segment of the past procedures to appraise the decisions in perspective of their relationship of value are Spearman [11], Analytical Hierarchical Process (AHP) [7], [13], [15] and Traveling Salesman Problem (TSP) [26]. The fragile information's among various components [19], [21], [28], [30], [32], [34], [36] among the bank stock model are managed by late secured systems [23], [25], [27], [29], [31], [33], [35], [37].

In boxplot method, the input data set is split to quartiles. In a boxplot, it has a minimum value, lower quartile, median, upper quartile, maximum value. Boxplot, it contains one box, it goes from lower quartile to upper quartile. The difference between upper quartile and lower quartile is the length of the box. Inside the box of boxplot, one vertical line is drawn, it is the median of the dataset. Median of the lower samples is called “Lower quartile” and Median of the higher samples is called “Upper quartile”. In the outside of the box in a boxplot, two more vertical lines are drawn, one vertical near upper quartile is called upper whisker and another one line near lower quartile is called lower whisker is shown in the following Fig. 1. The easiest way to find the quartiles have first sorted the data and take the minimum and maximum values as lower bound and upper bound respectively. Lower quartile, median upper quartile is we can find using the following methods in Section 2.

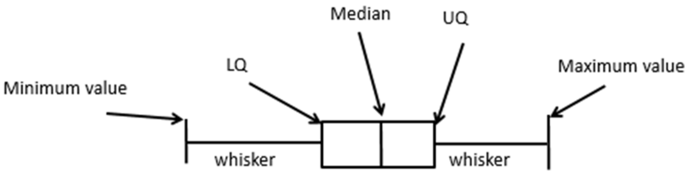

Fig. 1. Box Plot Attributes.

II. Data Analysis

A. Box Plot

Step 1: Sort the data on a primary attribute.
Step 2: Calculate the Median.
Step 3: Calculate the Quartiles.

Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)

Inter-quartile range: $IQR = Q_3 - Q_1$

Five number summary: min, Q_1, M, Q_3, max

Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually

Step 4: Calculate the Outlier:
More than 1.5 x IQR.

III. Calculation and Discussions

This is the sample dataset that we are going to know how the boxplot method works.

TABLE III. SAMPLE DATASET OF 1ST PART BETWEEN AGE 25 TO 45.

Age	Smoking status	Years smoked	Average per day	Gender	Grade
25	Smoker	12	15	Male	Nil
21	Non Smoker	0	0	Male	Nil
22	Former Smoker	5	2	Male	Nil
28	Smoker	10	8	Female	PG
35	Smoker	7	3	Male	PG
18	Former Smoker	8	2	Female	PG
19	Non Smoker	0	0	Female	PG
40	Smoker	12	6	Male	PG
45	Smoker	45	4	Female	PG
23	Smoker	2	5	Male	PG

TABLE IV. SAMPLE DATASET OF 2ND PART BETWEEN AGE 25 TO 45.

Race	Height	Weight	Family history	Cancer	Year
Asian	180	75	Yes	Yes	2005
Asian	178	80	No	No	2004
Asian	165	78	No	No	2005
Asian	178	79	Yes	No	2004
Asian	189	75	Yes	Yes	2003
Asian	175	80	Yes	No	2005
Asian	148	79	No	Yes	2005
Asian	168	72	Yes	No	2003
Asian	189	85	No	No	2004
Asian	168	69	No	No	2005

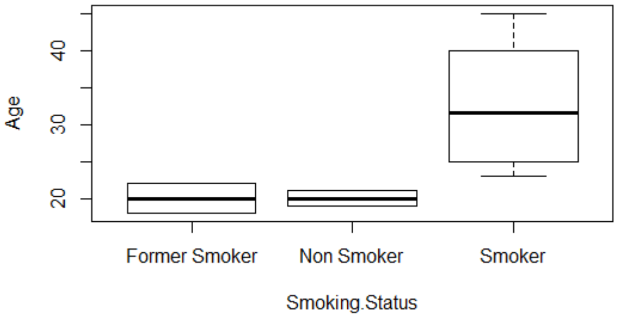

Fig. 2. Smokers by Ages.

The above boxplot shows that when comparing to former smoker and nonsmoker, the smoker is having higher chances of getting affected by lung cancer, from the boxplot of a smoker having a higher median, when comparing to age attribute from people having age 25 to 40 are high chances for cancer disease.

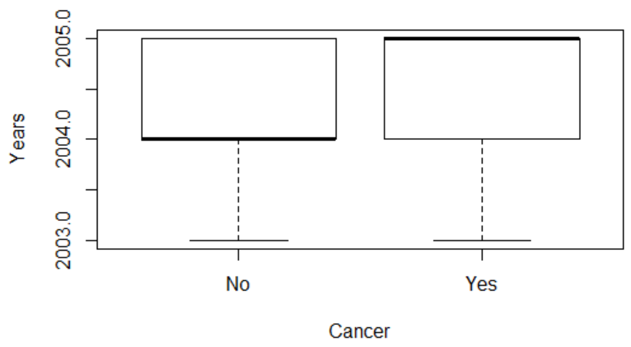

Fig. 3. Cancer in Years (2003 – 2005).

Above boxplot shows that when comparing the years 2004 to 2005, in year the boxplot for getting affected by cancer the chances is very low, because the median is in the lower quartile and people in 2005, having higher chances of getting cancer disease, because the median is near the upper quartile, we can understand this easily from the boxplot.

IV. Numerical Result Analysis

A. Boxplot for cancer in years

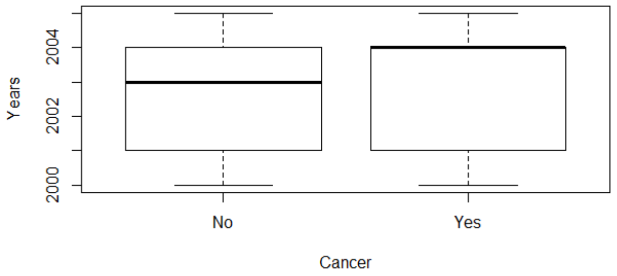

Fig. 4. Box plot for Cancer in Years.

In the above Fig. 4, from the median, we can easily understand that the number of people affected by cancer is increased with comparing to a nonsmoker.

B. Boxplot for all the attributes in the dataset

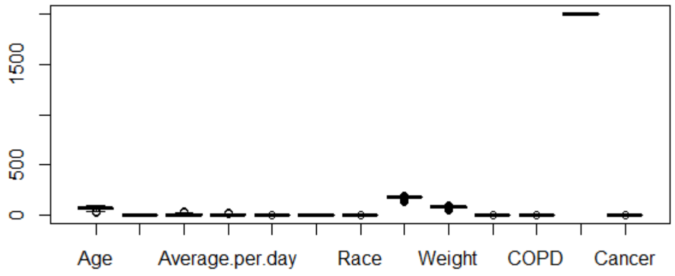

Fig. 5. Boxplot for Overall Attributes.

In the above Fig. 5 shows boxplot for all attributes with outliers.

C. Boxplot for Smoking status based on Age

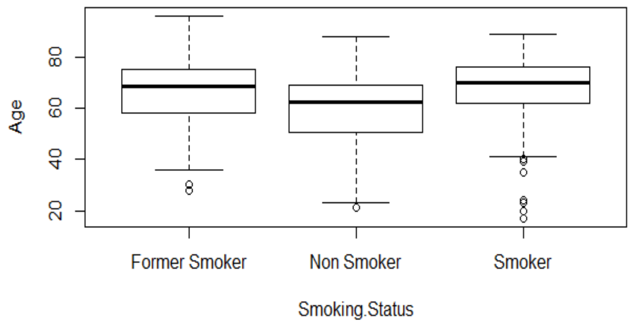

Fig. 6. Boxplot for Smokers by Ages.

From the above Fig. 6, it shows that the age between 60 to 80, people those who are smokers and former smoker are having higher chances to get cancer with comparing to a nonsmoker.

D. Boxplot for Smoking status based on Year

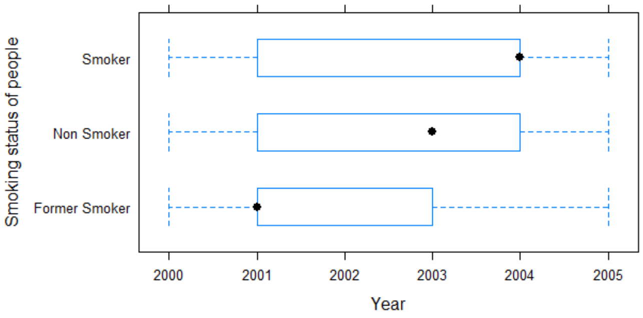

Fig. 7. Boxplot for Smoking Status of the Peoples (2000 – 2005).

The numbers of smokers are increased in 2004 when compared to the year 2000 – 2005. Former smokers also having fewer chances of getting lung cancer disease with compared to nonsmoker and smoker.

E. GG plot for Smoking Status vs Years Smoked

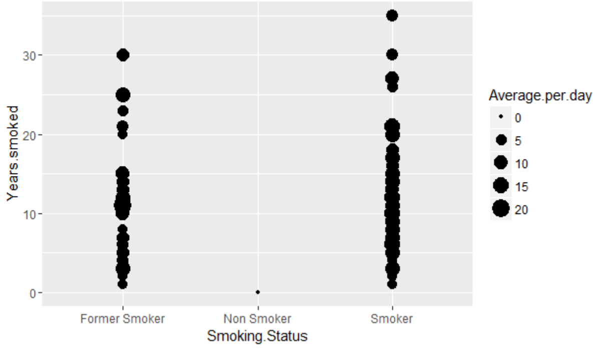

Fig. 8. GG plot for Smoking Status vs Years Smoked.

In Fig. 8 shows the average numbers of cigarette smokers are high when compared to former smoker and nonsmoker. Here, a maximum average of cigarette consumers per day is 20 and least is 0.

F. 3D plot for Lung Cancer

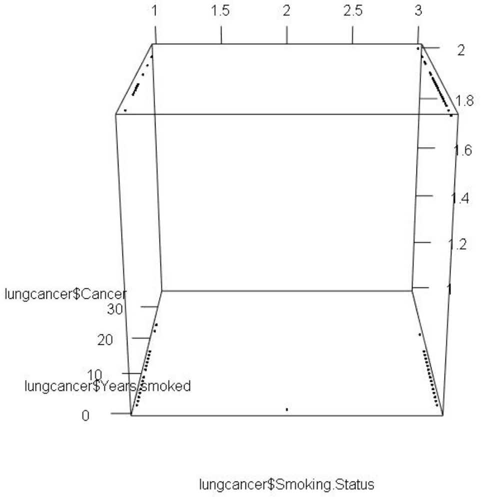

Fig. 9. The 3D plot of Lung Cancer.

Fig 9 shows the cancer, years smoked and smoking status.

G. Scatterplot for Lung cancer

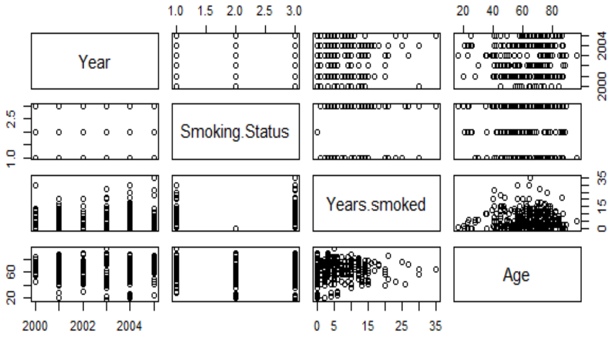

Fig. 10. Scatter plot of Year, Smoking Status, Years Smoked, and Age.

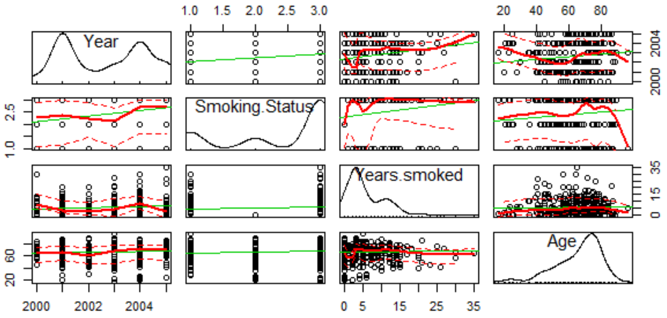

Fig. 11. Lung Cancer Causes Options.

From the above Fig. 11, scatterplot diagram we can easily make the relationship between the attributes. Here we have four attributes and four columns. The above scatterplot diagram first column for years, the second column for smoking status, and the third column for year's smoked and fourth one for age.

H. 3D Scatterplot

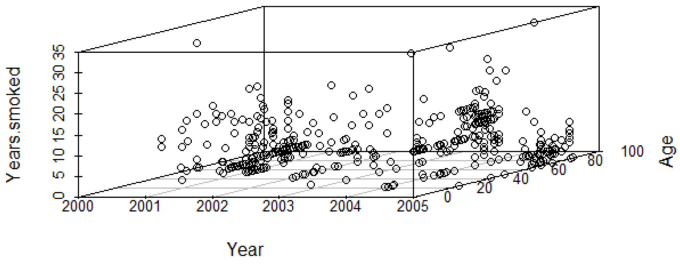

Fig. 12. 3D Scatter plot of Year, Years Smoked, and Age.

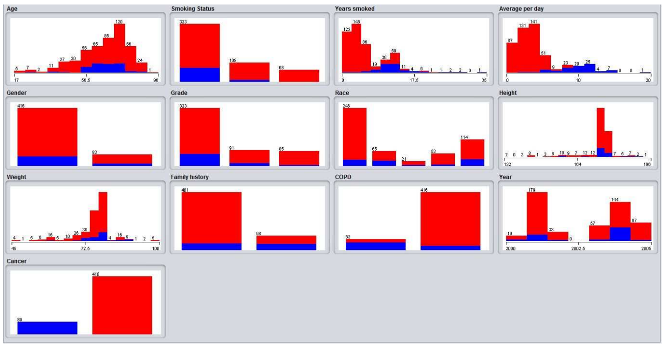

Fig. 13. Lung Cancer Chances.

In the above Fig 13 shows the getting chance for lung cancer for all the attributes in the datasets. In the above Fig. 13 first one age, shows that when the age between 55 to 90, this aged people who are having smoking habits have high chances of lung cancer disease.

I. Control chart

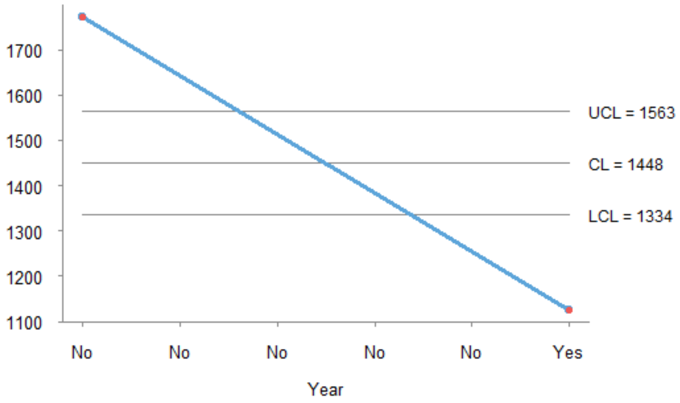

Fig. 14. C Chart for Cancer over a period of Years.

From the above Fig. 14, the upper control limit for age is 1563 (15.63), control limit is 1448(14.48) and the lower control limit is 1334(13.34). Cancer disease symptoms we can mostly identify between the age 13 to 15.

J. Control Chart

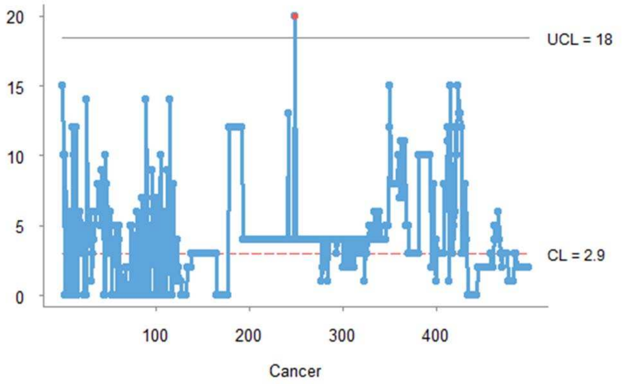

Fig. 15. C Chart of 300 Samples of Smokers for Lung Cancer Cause.

From the above control chart, upper control limit is 18 and control limit is 2.9. It is the control chart for all the data in the dataset. In the dataset of 300 samples, someone have high chances of getting lung cancer disease.

V. Conclusion

The purpose of this paper is to use “box and whisker plot” method for visualizing the samples of the dataset and from that results we can easily make relationships between the attributes. From the above boxplot method, we learned about which age of people mostly smoking people or farmer smoking people will have chances of getting lung cancer disease. we got some result with the help of these boxplot method results, we can make a system that gets some input from the user, so that can predicate whether the person has any chances to get cancer disease.

References

[1] Kampstra, Peter. "Beanplot: A boxplot alternative for visual comparison of distributions." Journal of statistical software 28, no. 1 (2008): 1-9.

[2] Frigge, Michael, David C. Hoaglin, and Boris Iglewicz. "Some implementations of the boxplot." The American Statistician 43, no. 1 (1989): 50-54.

[3] Benjamini, Yoav. "Opening the Box of a Boxplot." The American Statistician 42, no. 4 (1988): 257-262.

[4] Hubert, Mia, and Ellen Vandervieren. "An adjusted boxplot for skewed distributions." Computational statistics & data analysis 52, no. 12 (2008): 5186-5201.

[5] Thriumani, Reena, et al. "Cancer detection using an electronic nose: A preliminary study on detection and discrimination of cancerous cells." Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on. IEEE, 2014.

[6] Hauke J., Kossowski T., Comparison of values of Pearson’s and Spearman’s correlation coefficient on the same sets of data. Quaestiones Geographicae 30(2), Bogucki Wydawnictwo Naukowe, Poznań 2011, pp. 87–93, 3 figs, 1 table. DOI 10.2478/v10117-011-0021-1, ISBN 978-83-62662-62-3, ISSN 0137-477X.

[7] Piovani J.I., 2008. The historical construction of correlation as a conceptual and operative instrument for empirical research. Quality & Quantity 42: 757–777.

[8] P. Dhavachelvan, Chandra Segar T, K. Satheskumar, "Evaluation of SOA Complexity Metrics Using Weyuker’s Axioms," IEEE International Advance Computing (IACC), India, pp. 2325 – 2329, March 2009

[9] Halstead Metric for Intelligence, Effort, Time predictions, DOI:10.13140/RG.2.2.17988.42881

[10] Fisher R.A., 1921. On the “probable error” of a coefficient of correlation deduced from a small sample. Metron 1: 3–32.

[11] Spearman C.E, 1904b. General intelligence, objectively determined and measured. American Journal of Psychology 15: 201–293.

[12] Software metric Numerical Data analysis using Box plot and control chart methods, VIT University, DOI:10.13140/RG.2.2.27422.95041

[13] Vaishnavi B, Karthikeyan J, Kiran Yarrakula, Chandrasegar Thirumalai, “An Assessment Framework for Precipitation Decision Making Using AHP”, International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[14] Griffith D.A., 2003. Spatial autocorrelation and spatial filtering. Springer, Berlin.

[15] Chandrasegar Thirumalai, Senthilkumar M, “An Assessment Framework of Intuitionistic Fuzzy Network for C2B Decision Making”, International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[16] Rodgers J.L. & Nicewander W.A., 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42 (1): 59–66.

[17] F. Fioravanti, P. Nesi, “A method and tool for assessing object-oriented projects and metrics management,” Journal of Systems and Software, Volume 53, Issue 2, 31 August 2000, Pages 111-136

[18] Galton F., 1875. Statistics by intercomparison. Philosophical Magazine 49: 33–46

[19] Chandrasegar Thirumalai, Viswanathan P, “Diophantine based Asymmetric Cryptomata for Cloud Confidentiality and Blind Signature applications,” JISA, Elsevier, 2017.

[20] Galton F., 1877. Typical laws of heredity. Proceedings of the Royal Institution 8: 282–301.

[21] Chandrasegar Thirumalai, Sathish Shanmugam, “Multi-key distribution scheme using Diophantine form for secure IoT communications,” IEEE IPACT 2017.

[22] Galton F., 1888. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London 45: 135–145.

[23] Chandrasegar Thirumalai, Senthilkumar M, “Spanning Tree approach for Error Detection and Correction,” IJPT, Vol. 8, Issue No. 4, Dec-2016, pp. 5009-5020.

[24] Galton F., 1890. Kinship and correlation. North American Review 150: 419–431.

[25] Chandrasegar Thirumalai, Senthilkumar M, “Secured E-Mail System using Base 128 Encoding Scheme,” International journal of pharmacy and technology, Vol. 8 Issue 4, Dec. 2016, pp. 21797-21806.

[26] M.Senthilkumar, T.Chandrasegar, M.K. Nallakaruppan, S.Prasanna, “A Modified and Efficient Genetic Algorithm to Address a Travelling Salesman Problem,” in International Journal of Applied Engineering Research, Vol. 9 No. 10, 2014, pp. 1279-1288

[27] Nallakaruppan, M.K., Senthil Kumar, M., Chandrasegar, T., Suraj, K.A., Magesh, G., “Accident avoidance in railway tracks using Adhoc wireless networks,” 2014, IJAER, 9 (21), pp. 9551-9556.

[28] T Chandra Segar, R Vijayaragavan, “Pell's RSA key generation and its security analysis,” in Computing, Communications and Networking Technologies (ICCCNT) 2013, pp. 1-5

[29] Chandrasegar Thirumalai, Senthilkumar M, Vaishnavi B, “Physicians Medicament using Linear Public Key Crypto System,” in International conference on Electrical, Electronics, and Optimization Techniques, ICEEOT, IEEE & 978-1-4673-9939-5, March 2016.

[30] Chandrasegar Thirumalai, “Physicians Drug encoding system using an Efficient and Secured Linear Public Key Cryptosystem (ESLPKC),” International journal of pharmacy and technology, Vol. 8 Issue 3, Sep. 2016, pp. 16296-16303

[31] E Malathy, Chandra Segar Thirumalai, "Review on non-linear set associative cache design," IJPT, Dec-2016, Vol. 8, Issue No.4, pp. 5320-5330

[32] “DDoS: Survey Of Traceback Methods”, International Joint Journal Conference in Engineering 2009, ISSN 1797-9617.

[33] Chandrasegar Thirumalai, Senthilkumar M, Silambarasan R, Carlos Becker Westphall, “Analyzing the strength of Pell’s RSA,” IJPT, Vol. 8 Issue 4, Dec. 2016, pp. 21869-21874.

[34] Chandramowliswaran N, Srinivasan.S and Chandra Segar T, “A Novel scheme for Secured Associative Mapping” The International J. of Computer Science and Applications (TIJCSA) & India, TIJCSA Publishers & 2278-1080, Vol. 1, No 5 / pp. 1-7 / July 2012

[35] Chandrasegar Thirumalai, “Review on the memory efficient RSA variants,” International Journal of Pharmacy and Technology, Vol. 8 Issue 4, Dec. 2016, pp. 4907-4916.

[36] Vinothini S, Chandra Segar Thirumalai, Vijayaragavan R, Senthil Kumar M, “A Cubic based Set Associative Cache encoded mapping,” International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 02 May -2015

[37] Chandrasegar Thirumalai, Himanshu Kar, “Memory Efficient Multi Key (MEMK) generation scheme for secure transportation of sensitive data over Cloud and IoT devices,” IEEE IPACT 2017.