Data analysis using Box and Whisker Plot for Lung Cancer

Conference Paper · April 2017

DOI: 10.1109/IPACT.2017.8245071

3 authors, including:

Chandra Segar Thirumalai
VIT University
63 PUBLICATIONS **1,370** CITATIONS
SEE PROFILE

Vignesh Manickam
VIT University
1 PUBLICATION **14** CITATIONS
SEE PROFILE

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

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

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

upper quartile is we can find using the following methods in Section 2.

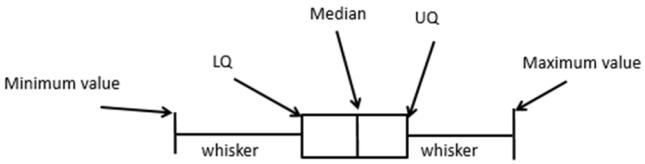

Fig. 1. Box Plot Attributes.

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 (25^{th} percentile), Q_3 (75^{th} 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

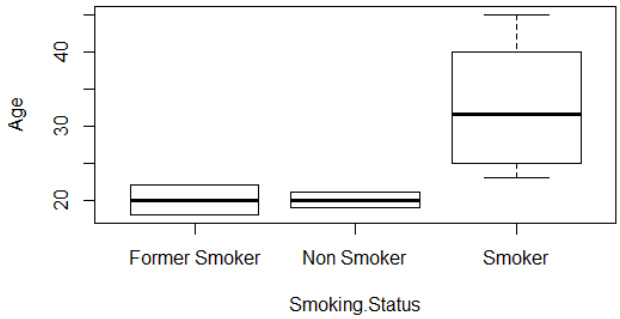

Fig. 2. Smokers by Ages.

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.

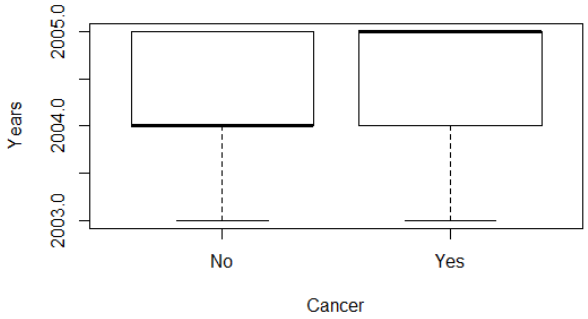

Fig. 3. Cancer in Years (2003 – 2005).

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

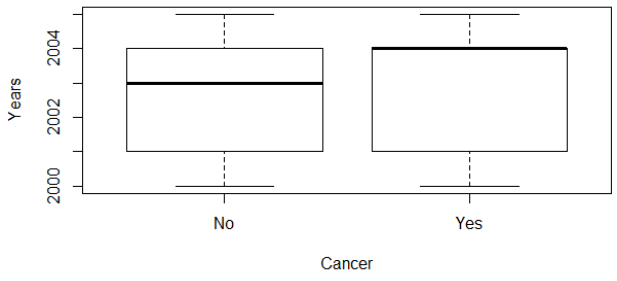

Fig. 4. Box plot for Cancer in Years.

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

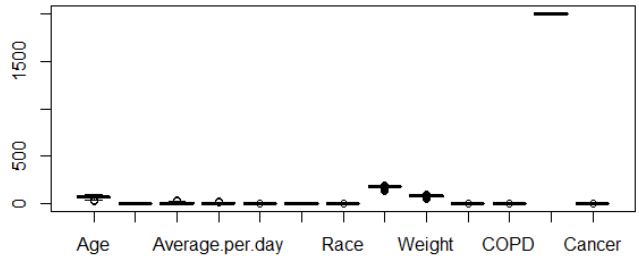

Fig. 5. Boxplot for Overall Attributes.

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

C. Boxplot for Smoking status based on Age

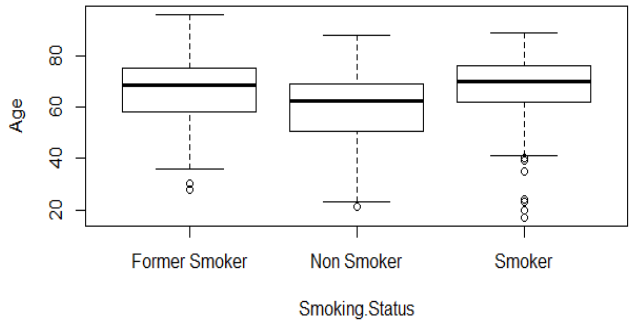

Fig. 6. Boxplot for Smokers by Ages.

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

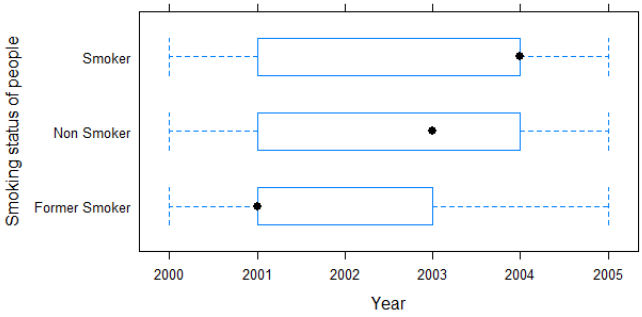

Fig. 7. Boxplot for Smoking Status of the Peoples (2000 – 2005).

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

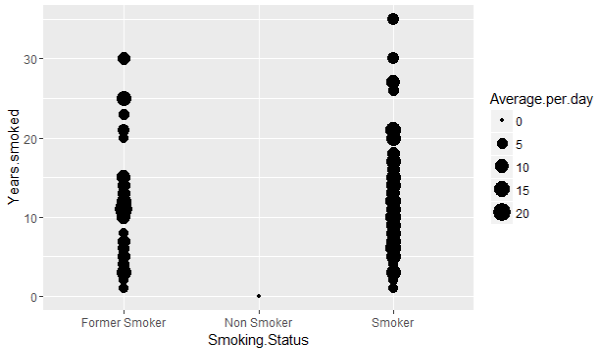

Fig. 8. GG plot for Smoking Status vs Years Smoked.

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

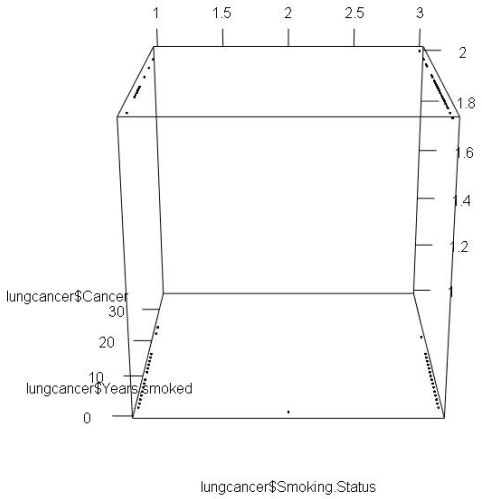

Fig. 9. The 3D plot of Lung Cancer.

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

G. Scatterplot for Lung cancer

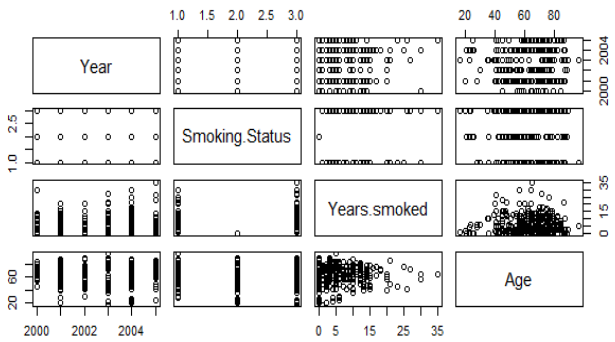

Fig. 10. Scatter plot of Year, Smoking Status, Years Smoked, and Age.

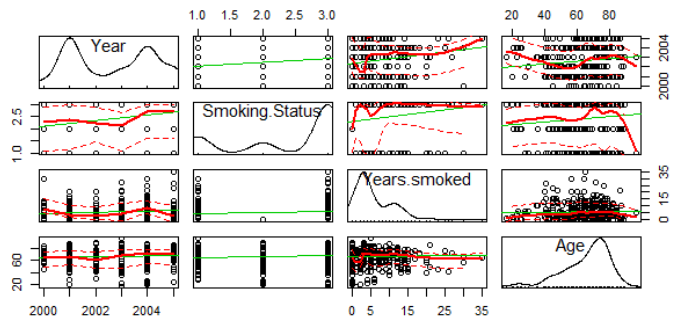

Fig. 11. Lung Cancer Causes Options.

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

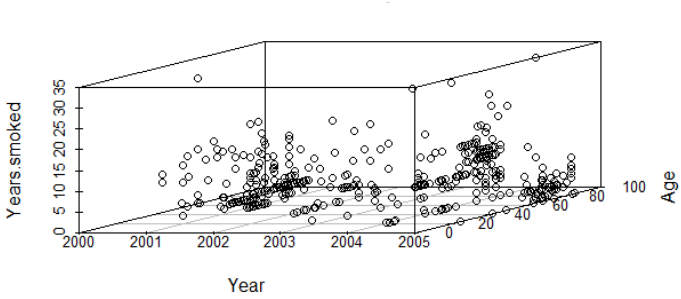

Fig. 12. 3D Scatter plot of Year, Years Smoked, and Age.

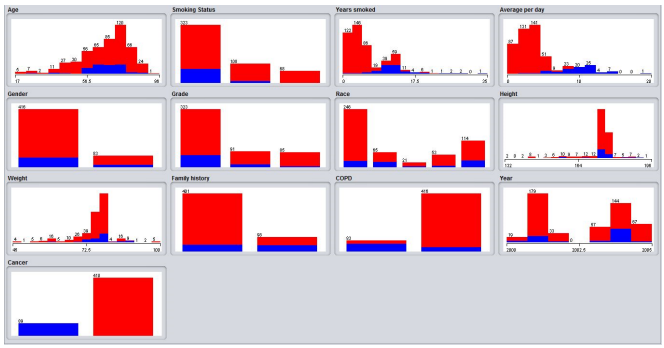

Fig. 13. Lung Cancer Chances.

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

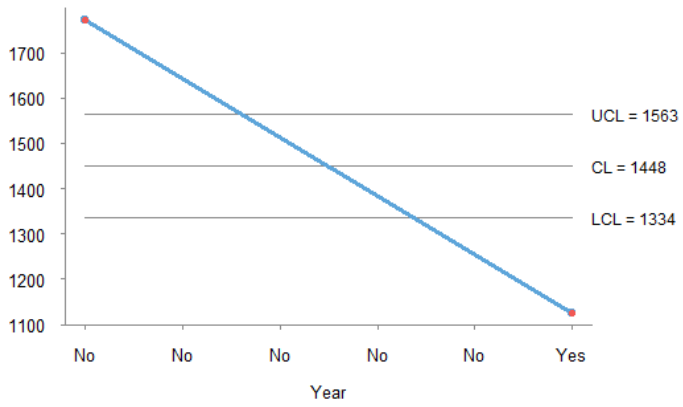

Fig. 14. C Chart for Cancer over a period of Years.

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

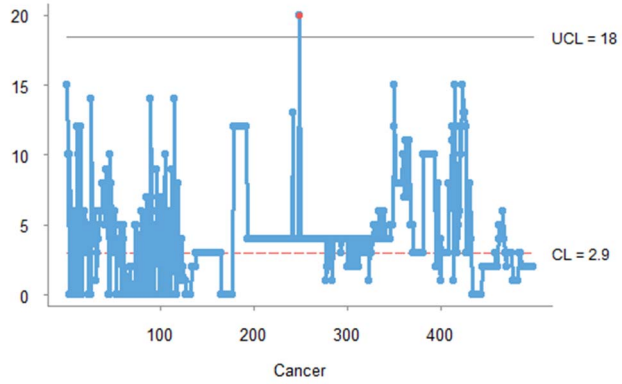

Fig. 15. C Chart of 300 Samples of Smokers for Lung Cancer Cause.

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] 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.

[6] 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.

[7] 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

[8] 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

[9] 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

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

[11] Chandrasegar Thirumalai, Rashad Manzoor, "Cost Optimization using Normal Linear Regression Method for Breast Cancer Type I Skin," IEEE IPACT 2017.

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

[13] 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

[14] McWilliams, Annette, et al. "Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose." IEEE Transactions on Biomedical Engineering 62.8 (2015): 2044-2054.

[15] 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

[16] Bromis, Konstantinos, et al. "Analysis of resting state and task-related fMRI data in small cell lung cancer patients before undertaking PCI." Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on. IEEE, 2014.

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

[18] Dharmarajan, A., and T. Velmurugan. "Lung cancer data analysis by k-means and farthest first clustering algorithms." Indian Journal of Science and Technology 8.15 (2015).

[19] 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.

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

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

[22] Avinash, S., K. Manjunath, and S. Senthil Kumar. "An improved image processing analysis for the detection of lung cancer using Gabor filters and watershed segmentation technique." Inventive Computation Technologies (ICICT), International Conference on. Vol. 3. IEEE, 2016.

[23] 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.

[24] Zhang, Wenbin, Jian Tang, and Nuo Wang. "Using the machine learning approach to predict patient survival from high-dimensional survival data." Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. IEEE, 2016.

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

[26] Rodriguez, Rafael, et al. "Low-order statistical analysis of 1-D diffuse reflectance signals from cancer cells using 2-D scalogram images." Imaging Systems and Techniques (IST), 2016 IEEE International Conference on. IEEE, 2016.

[27] 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

[28] Chandramowliswaran N, Srinivasan.S, and Chandra Segar.T, "A Note on Linear based Set Associative Cache address System" International J. of Computer Science and Engg. (IJCSE) & India, Engineering Journals & 0975-3397, Vol. 4 No. 08 / pp. 1383-1386 / Aug. 2012.

[29] Benezi, Sofia, et al. "Tract-Based Spatial Statistics analysis of diffusion-tensor imaging data in patients with Small Cell Lung Cancer." Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on. IEEE, 2014.

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

[31] Vinothini S, Chandra Segar Thirumalai, Vijayaragavan R, "Analyzing the performance of AFRA with its traditional routing," IRJET, Vol. 2 No. 2, May 2015, pp.373-382

[32] Anderson, Paul E., et al. "Predictive modeling of lung cancer recurrence using alternative splicing events versus differential expression data." Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on. IEEE, 2014.

[33] 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.

[34] Amudhavel, J., et al. "Effective maintenance of replica in distributed network environment using DST." Advances in Recent Technologies in Communication and Computing (ARTCom), 2010 International Conference on. IEEE, 2010.

Chandra Segar Thirumalai was born in Pondicherry capital city, Indian union territory of Puducherry, in 1983. He received the Bachelor of Engineering in computer science and engineering from Dr. Paul's Engineering College, Anna University, Chennai, India, in 2005 and Master of Technology in computer science and engineering from Pondicherry Central University, Pondicherry, India, in 2009. He is currently doing his Doctorate of Philosophy in School of Information Technology and Engineering at VIT University, Vellore, India.

He has been working as Assistant Professor Senior in the Department of Digital Communications, School of Information Technology and Engineering at VIT University, Vellore, India. His area of specialization includes Public Key Cryptography, and Networking. He is the author of more than fifteen International journals and five IEEE conferences. He is a professional member of IEEE and IDES. He received the GATE score conducted by MHRD, India on 2009 with 87.28% in open general cadre and also qualified in SET conducted by Tamil Nadu, India on 2016. He also received VIT Most Active Researcher Award from the year 2011 to 2016.

Vignesh M Currently pursuing MS Software Engineering at VIT University, School of Information Technology, Vellore, India.

Balaji R Currently pursuing MS Software Engineering at VIT University, School of Information Technology, Vellore, India.