International Journal of Emerging Technology and Innovative Engineering
Volume 5, Issue 2, February 2019 (ISSN: 2394 – 6598)

CHURN PREDICTION IN TELECOM USING CLASSIFICATION ALGORITHMS

G.M. Apurva Sree[1], S. Ashika[1], S. Karthi[1],V. Sathesh[1], M. Shankar[1], J. Pamina[2]

[1]PG Students, Sri Krishna College of Technology, Coimbatore,

[2]Assistant Professor, Sri Krishna College of Technology, Coimbatore,

ABSTRACT

The term churn is said to be when customers or people move from one telecom service provider to another. Churn prediction is the process of predicting whether there is a chance for any customer or people to change from one telecom service provider to another. In recent times, the problem can be predicted using advanced algorithms like support vector machine, logistic regression, random forest algorithm. Churning rate can also be analyzed by several case machine learning algorithms. In this paper, we have summarized a comparative study on rate of churning of customers using different algorithms.

KEYWORDS:

Random forest approach, Logistic Regression, Support Vector Machine, Data Visualization

1. INTRODUCTION:

Customers have a wide range of telecom services and they choose among the best possible services from the telecom industry and switching from one service to other is called churn[2], increased customer churn is always a major concern today[3]. The accuracy rate also makes us to know about the customers who are not willing to switch over to the other telecom using churn prediction [4]. Logistic regression research indicates that the modern churn prediction is also possible with the help of clustering algorithm [5]. Here the factors affecting the churn prediction rate for the given telecom service is given with the algorithms namely:

- Random forest
- Logistic regression
- SVM

2. DATA PREPROCESSING:

2.1. EXPLORATORY DATA ANALYTICS:

Exploratory Data Analytics is the way of visualizing the output from the data's in a understandable way, here the basic use of data visualization is mainly to help and analyze out the characteristic features of each separate data sets together. Also the charts, comparative analysis have been very well developed with the calculation by using the Exploratory Data

Analytics thus makes us to predict the behavior of churn and makes us to apply the various methodologies.

2.1.1. PSEUDOCODE FOR EDA:

$D_0 \leftarrow$ Generate M individuals (the initial population) randomly

Repeat for $l = 1, 2, \ldots$ until a stopping criterion is met

$D_{l-1}^N \leftarrow$ Select $N \leq M$ individuals from D_{l-1} according to a selection method

$\rho_l(\boldsymbol{x}) = \rho(\boldsymbol{x}|D_{l-1}^N) \leftarrow$ Estimate the probability distribution of an individual being among the selected individuals

$D_l \leftarrow$ Sample M individuals (the new population) from $\rho_l(\boldsymbol{x})$

From the given IBM WATSON dataset, month to month contracts, absence of online security and tech support seem to be positively correlated with churn. While, tenure, two year contracts seem to be negatively correlated with churn. Interestingly, services such as online security, streaming TV, online backup, tech support, etc. without internet connection seem to be negatively related to churn. We will explore the patterns for the above correlations below before we delve into modeling and identifying the important variables.

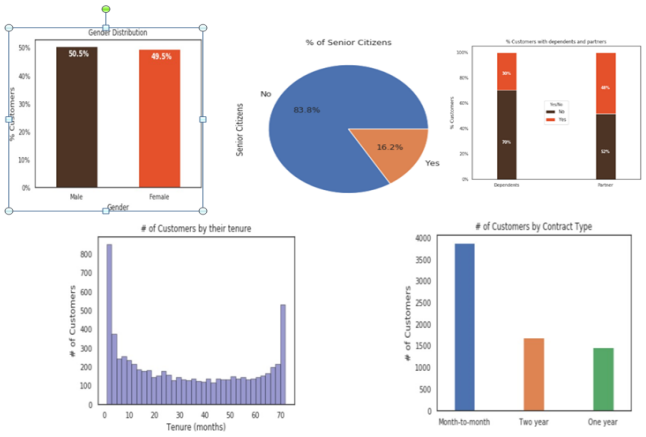

Fig. 2.1 shows the gender distribution, percentage of senior citizens, percentage of customers with dependents and partners, customers by tenure and contract type.

The above mentioned histograms gives us a detailed understanding about the distribution rates according with the gender and the population of senior citizens about 16.2% people who are under senior citizen category and 83.8 % of people are younger. Thus it gives us the clarity of understanding about the different age groups. We can analyze that among the customers who have a partner, only about half of them also have a dependent, while other half do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of

them do not have any dependents. So we can say that about 30% of total customers are dependent while 50% have partners. It is clear from the above observation. It is known that most of the customers are under month – month contracts, and where there are equal no of customers in both two year and one year contract.

2.2. TENURE OF CUSTOMERS BASED ON CONTRACTS:

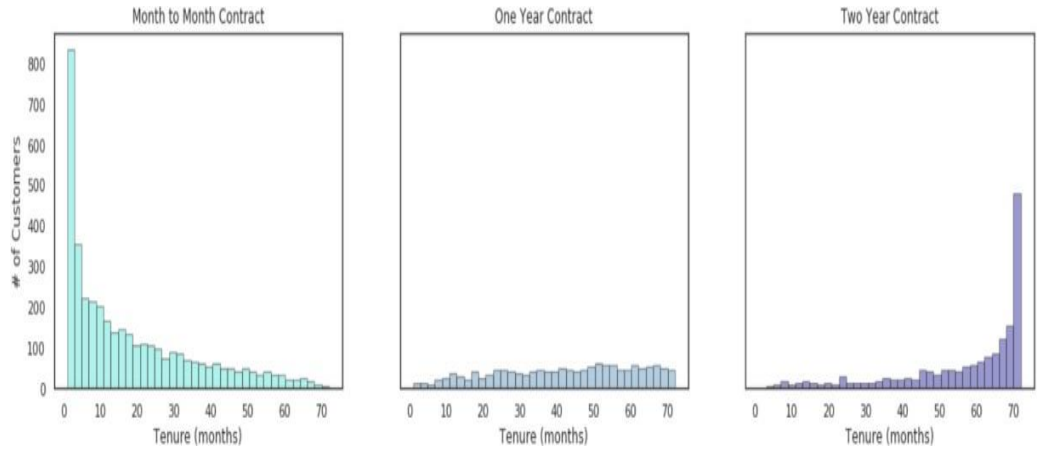

Fig 2.2 Showing customers contract in terms of Tenure

From the above mentioned histogram it is clear that, most of the monthly contracts last for 1- 2 months, while the 2 year contracts tend to last for about 70 months. This shows that the customers taking a longer contract are more loyal to the company and tend to stay with it for a longer period of time.

2.3. DISTRIBUTION OF VARIOUS SERVICES USED BY CUSTOMER:

From the given histogram below, we have an idea of the various services used by customers they are phone service, online security, tech support, multiple lines, online backup, streaming TV, internet service, device protection, streaming movies.

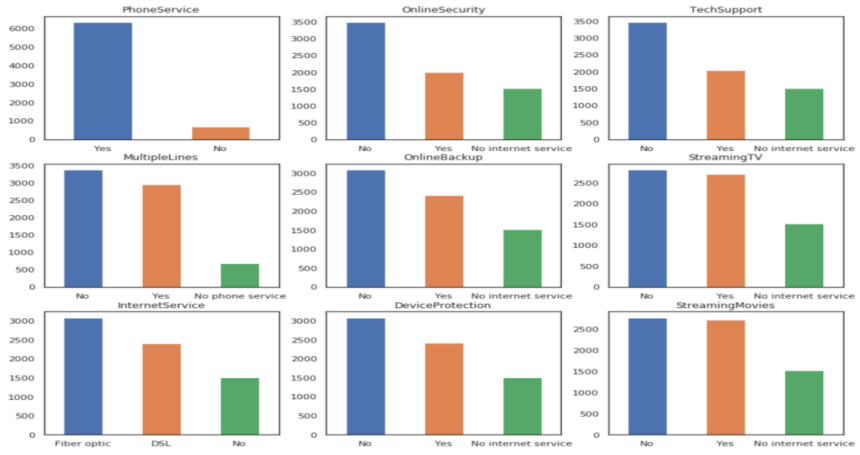

Fig 2.3 showing details on the various services offered to the customers

2.4. RELATIONSHIP BETWEEN MONTHLY AND TOTAL CHARGES:

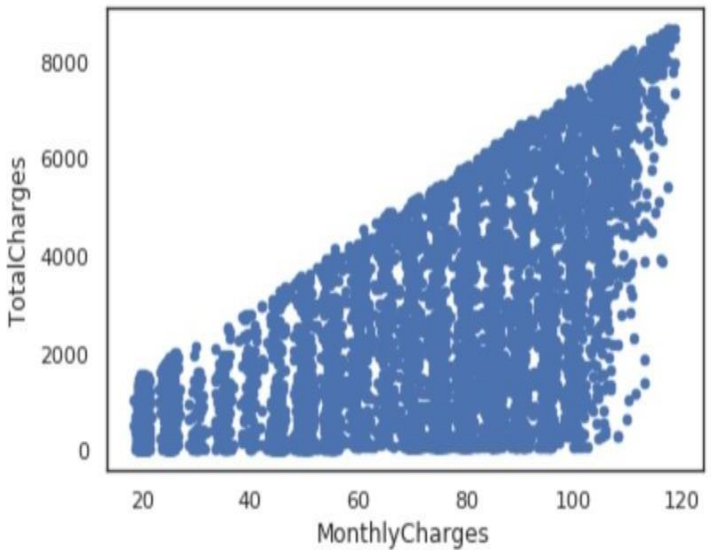

Fig 2.4 Applying hyper lanes for the given two attributes

2.5.INTERACTION OF CHURN WITH OTHER VARIABLES:

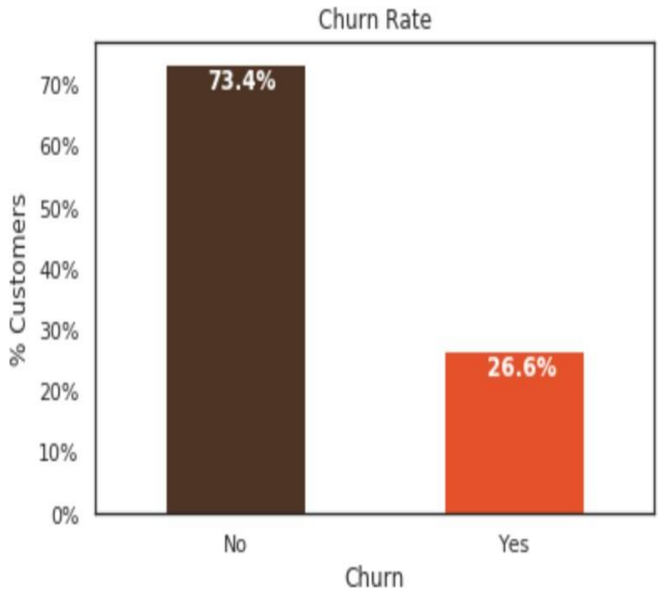

Fig. 2.5.1 shows the percentage of churn rate among customers

From the taken data set, it is known that 26.6% of customers are churned from one service provider to another, while 73.4 % of customers do not churn.

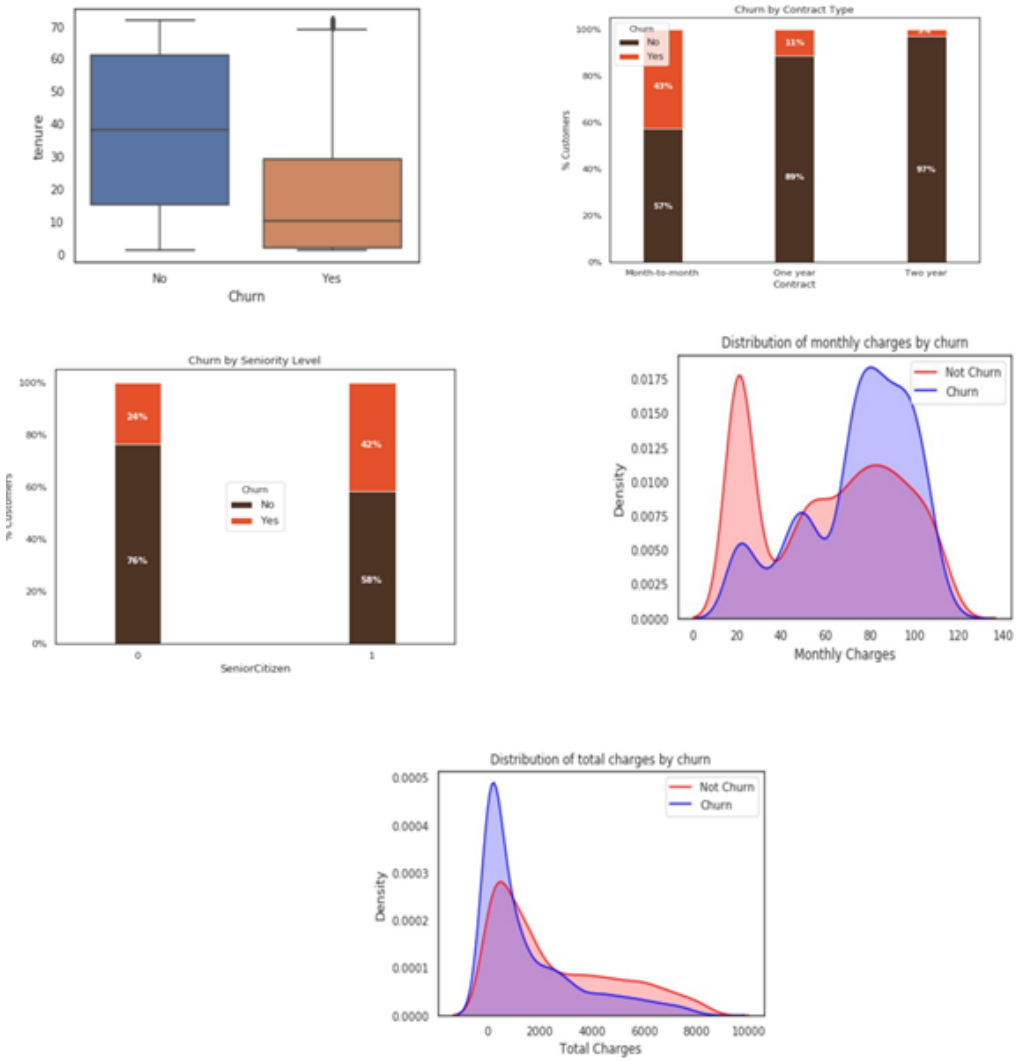

Fig. 2.5.2. shows the interaction of churn vs. tenure, Churn by contract type, Churn by monthly charges, Churn by total charges

Customers who do not churn is expected to have long service with their previous telecom. The churn rate is very higher for customers who have a month to month contract. From the above plot, it is clear that the senior citizens churn more than younger people. By the above graph, it is clear that when the monthly charges are high, people churn more. The above graph shows that there is higher churn when the total charges are lower.

3. ALGORITHMS :

3.1. SUPPORT VECTOR MACHINE:

SVM is the supervised learning model that analyses the given data used for the purpose of classification and regression approach, mainly here the data sets from the churn prediction are analyzed with the total dependency between the data sets for the churned data .SVM plays a part by listing the data sets for calculating the classification and regression analysis

3.1.1. PSEUDOCODE FOR EXECUTING SVM:

Algorithm SVM-RCE-EC (input data *D*)
X= the training dataset
s = genes list (all the genes) or top *n_g* genes by t-test
n = infinity number
m = final number of clusters
d = the reduction parameter
While ($n >= m$) do

1. n= Cluster the given genes S into clusters $s_1, s_2, \ldots, s_n$ using Ensemble Clustering (**EC Clustering step**)
2. For each cluster $i=1..n$ calculate its $Score(X(s_i), f, r)$ (**SVM scoring step**)
3. Remove the $d\%$ clusters with lowest score (**RCE step**)
4. Merge surviving genes again into one pool S

3.2. RANDOM FOREST ALGORITHM:

Random forest algorithm approach is a way of collection of different decision trees together without the process of disturbing the classification and prediction approaches being applied in the churn prediction process being a large number of data sets available within this process this random forest algorithm is useful. Random forest algorithm is implemented out here by using the multiple level of the decision trees and these decision trees are here constructed between the different factors of the churn prediction like :Telecom providers, Internet service, Call rates, Messaging services , etc. and the mode and the medians are calculated meanwhile consistency of the data sets are established and the data sets are calculated and the regression rates are very well established here. The details of different data analytics algorithms are analysed and the accuracy values are provided here.

3.2.1. PSEUDOCODE FOR RANDOM FOREST PRECONDITION:

A training set S := (x1, y1), . . . ,(xn, yn), features F, and number of trees in forest B. 1 function RandomForest(S , F)
2 H ← ∅

3 for i ∈ 1, . . . , B do

4 S (i) ← A bootstrap sample from S

5 hi ← Randomized TreeLearn(S (i) , F) 6 H ←
H ∪ {hi}

7 end for

8 return H

9 end function

10 function Randomized TreeLearn(S , F) 11 At
each node:
12 f ← very small subset of F 13
Split on best feature in f 14 return
The learned tree
15 nd function

3.3. LOGISTIC REGRESSION:

Logistic Regression method determines an outcome by analyzing a dataset with one or more independent variables. It is a statistical process which predicts the probability of an outcome that possesses only have two values (called as dichotomy). The logistic regression is similar to that of the linear regression. In multiple logistic regression model, the dependent variable is binary (dichotomous) where in the dataset it contains 1(true) and 0(false) in the data code. In linear regression, the values would be predicted outside the range of 0 and 1. Meanwhile the logistic regression produces curves in between the values 0 and 1.

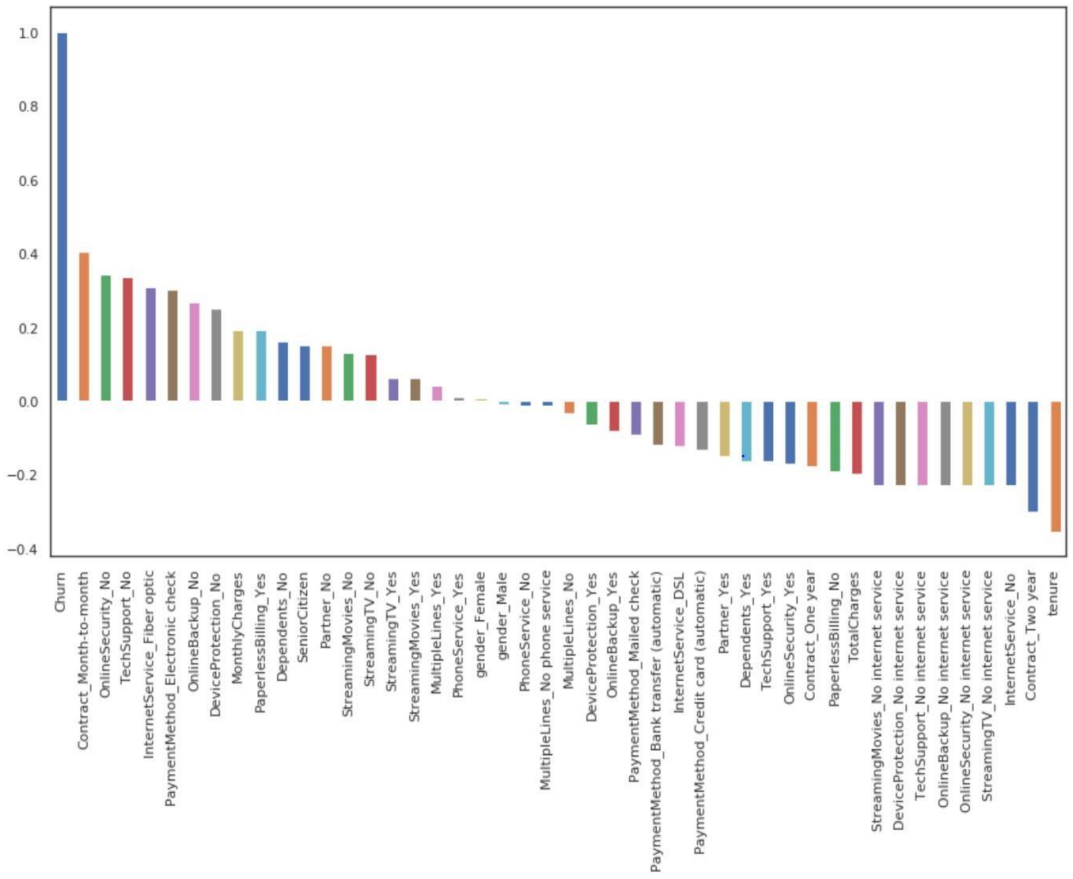

Fig 3.3. shows the churn rate using random forest, SVM, Logistic Regression

An analysis on the above datasets have been made using the data visualization, SVM and random forest algorithm the dependency of various variables between the datasets have been found for churn prediction. The dataset holds the details of about 7000 customer details approximately, which could be easily analyzed by random forest algorithm which can handle large number of datasets. By mapping the entities, the relationship between the factors influencing churn prediction are found. The different data sets use different algorithms and the way of solving heterogeneity, scalability always plays a major role in this accuracy process. Here the data sets being large in number always maintaining a correct accuracy always plays an important role.

4. ACCURACY :

ALGORITHMS	**ACCURACY**
Logistic Regression	80.75 %
Random Forest	80.88%
Support Vector Machine	**82%**

Table 4. represents accuracy calculation.
*Higher values are in bold.

5. CONCLUSION:

The churn prediction technique and the data algorithms, data analytics plays a major part in the present digital era as the data's gathered out from the various machine learning algorithms and the data analytics covers a wide scope helping us to know about the different factors affecting the churning rate also it can bring out the predictive capacity of the customers mind set enabling the various telecom service providers to change over to the new schemes possible, also it leaves us with the various innovative application of machine learning algorithms and the data analytics approach to solve out the present challenges facing the society also the various need for the data analysts and the importance of data's have been very well pictured out by the algorithms, This in turn also opens out a new gateway for the data analysts and the application of various innovative descriptive, predictive and prescriptive algorithms possible.

6. FUTURE WORKS:

Churn prediction and various factors causing the customers to switch over from one service provider to other circle is unavoidable and the above mentioned attributes and algorithms mentioned will be helpful in analyzing out the conditions causing the churn to occur. Churn prediction have wide applications in the near future making it an unavoidable component for the near future.

REFERENCES

[1] Idris,A., Iftikhar,A. & Rehman, Z.. Cluster Computer (2017). https://doi.org/10.1007/s10586-017-1154-3

[2] Azeem, M., Usman, M. & Fong, A.C.M. Telecommunication System (2017) 66: 603. https://doi.org/10.1007/s11235-017-0310-7

[3] Sivasankar, E. &Vijaya, J. Neural Computer & Application (2018). https://doi.org/10.1007/s00521-018-3548-4

[4] Egyptian Informatics Journal, ISSN: 1110-8665, Vol: 18, Issue: 3, Page: 215-220

[5] W. Bi, M. Cai, M. Liu and G. Li, "A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn," in *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1270-1281,June2016. doi: 10.1109/TII.2016.2547584

[6] N. Lu, H. Lin, J. Lu and G. Zhang, "A Customer Churn Prediction Model in Telecom Industry Using Boosting," in *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1659-1665, May 2014. doi: 10.1109/TII.2012.2224355.

[7] Deepa, V., A. Jenifa, and J. Pamina. "APPROACHES BASED ON DATA MINING IN NATURAL LANGUAGE PROCESSING."

[8] Raja, J. Beschi, S. Chenthur Pandian, and J. Pamina. "Certificate revocation mechanism in mobile ADHOC grid architecture." *Int. J. Comput. Sci. Trends Technol* 5 (2017): 125-130.

[9] Raja, J. Beschi, and K. Vivek Rabinson. "Iaas for Private and Public Cloud using Openstack." International Journal of Engineering 5.04 (2016).

[10] Raja, J. Beschi, and V. Vetriselvi. "Mobile Ad Hoc Grid Architecture Based On Mobility of Nodes." International Journal of Innovative Research in Computer and Communication Engineering 2 (2014): 49-55.

[11] Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2018). Financial crisis prediction model using ant colony optimization. International Journal of Information Management.

[12] Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2018). Intelligent hybrid model for financial crisis prediction using machine learning techniques. Information Systems and e-Business Management, 1-29.

[13] Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, 374-382.

[14] Lakshmanaprabu, S. K., Shankar, K., Gupta, D., Khanna, A., Rodrigues, J. J., Pinheiro, P. R., & de Albuquerque, V. H. C. (2018). Ranking analysis for online customer reviews of products using opinion mining with clustering. Complexity, 2018.

[15] Karthikeyan, K., Sunder, R., Shankar, K., Lakshmanaprabu, S. K., Vijayakumar, V., Elhoseny, M., & Manogaran, G. (2018). Energy consumption analysis of Virtual Machine migration in cloud using hybrid swarm optimization (ABC–BA). The Journal of Supercomputing, 1-17.

[16] Shankar K, Mohamed Elhoseny, Lakshmanaprabu S K, Ilayaraja M, Vidhyavathi RM, Mohamed A. Elsoud, Majid Alkhambashi. Optimal feature level fusion based ANFIS classifier for brain MRI image classification. Concurrency Computat Pract Exper. 2018;e4887.https://doi.org/10.1002/cpe.4887

[17] Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & de Albuquerque, V. H. C. (2018). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The Journal of Supercomputing, 1-16.

[18] Lakshmanaprabu SK, K. Shankar, Ashish Khanna, Deepak Gupta, Joel J. P. C. Rodrigues, Plácido R. Pinheiro, Victor Hugo C. de Albuquerque, "Effective Features to Classify Big Data using Social Internet of Things", IEEE Access, Volume.6, page(s):24196-24204, April 2018.

[19] Andino Maseleno, Alicia Y.C. Tang, Moamin A. Mahmoud, Marini Othman, Suntiaji Yudo Negoro, Soukaina Boukri, K. Shankar, Satria Abadi, Muhamad Muslihudin, "The Application of Decision Support System by Using Fuzzy Saw Method in Determining the Feasibility of Electrical Installations in Customer's House", International Journal of Pure and Applied Mathematics, Vol.119, No. 16, page(s): 4277-4286, July 2018.

[20] Muhammad Muslihudin, Risma Wanti, Hardono, Nurfaizal, K. Shankar, Ilayaraja M, Andino Maseleno, Fauzi, Dwi Rohmadi Mustofa, Muhammad Masrur, Siti Mukodimah, "Prediction of Layer Chicken Disease using Fuzzy Analytical Hierarcy Process", International Journal of Engineering & Technology, Volume. 7, Issue-2.26, page(s): 90- 94, June 2018.

[21] Eka Sugiyarti, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda, K. Shankar, Andino Maseleno, "Decision Support System of Scholarship Grantee Selection using Data Mining", International Journal of Pure and Applied Mathematics, Volume.119, No. 15, page(s): 2239-2249, June 2018.

[22] Tri Susilowati, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda, K. Shankar, Andino Maseleno, Anis Julia, Sucipto, "Determination of Scholarship Recipients using Simple Additive Weighting Method", International Journal of Pure and Applied Mathematics, Volume.119, No. 15, page(s): 2231-2238, June 2018.

[23] E. Laxmi Lydia, K. Shankar, M. Ilayaraja, K. Sathesh Kumar, "Technological Solutions for Health Care Protection and Services Through Internet Of Things(IoT)", International Journal of Pure and Applied Mathematics, Volume 118, No. 7, page(s) 277-283, February 2018.

[24] E. Laxmi Lydia, K. Shankar, J. Pamina, J. Beschi Raja, "Correlating NoSQL Databases With a Relational Database: Performance and Space", International Journal of Pure and Applied Mathematics, Volume 118, No. 7, page(s) 235-244, February 2018.

[25] K. Shankar. "Prediction of Most Risk Factors in Hepatitis Disease using Apriori Algorithm", Research Journal of Pharmaceutical, Biological and Chemical Sciences (ISSN: 0975-8585, Volume 8, No. 5, page(s): 477-484, 2017.

[26] Haidi Rao, Xianzhang Shi, Ahoussou Kouassi Rodrigue, Juanjuan Feng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan, LichuanGu, Feature selection based on artificial bee colony and gradient boosting decision tree, Applied Soft Computing, Volume 74, Pages 634-642, January 2019.

[27] Baofu Fang, Xiaoping Guo, Zaijun Wang, Yong Li, Mohamed Elhoseny, Xiaohui Yuan, Collaborative task assignment of interconnected, affective robots towards autonomous healthcare assistant, Future Generation Computer Systems, Volume 92, Pages 241-251, March 2019.

[28] Noura Metawaa, M. Kabir Hassana, and Mohamed Elhoseny, "Genetic algorithm based model for optimizing bank lending decisions", Expert Systems with Applications, Volume 80, Pages 75–82, 2017.

[29] Elhoseny, M., Shankar, K., Lakshmanaprabu, S. K., Maseleno, A., & Arunkumar, N. (2018). Hybrid optimization with cryptography encryption for medical image security in Internet of Things. Neural Computing and Applications, 1-15. https://doi.org/10.1007/s00521-018-3801-x

[30] Shankar, K., Elhoseny, M., Kumar, R. S., Lakshmanaprabu, S. K., & Yuan, X. (2018). Secret image sharing scheme with encrypted shadow images using optimal homomorphic encryption technique. Journal of Ambient Intelligence and Humanized Computing, 1-13. https://doi.org/10.1007/s12652-018-1161-0

[31] K. Shankar, Mohamed Elhoseny, E. Dhiravida chelvi, SK. Lakshmanaprabu, Wanqing Wu, , IEEE Access, Vol.6, Issue.1, page(s): 77145-77154, December 2018. https://doi.org/10.1109/ACCESS.2018.2874026

[32] Muthukumar Murugesan, Dr K. Karthikeyan. "Business intelligence market trends and growth in enterprise business." International Journal on Recent and Innovation Trends in Computing and Communication 4.3 (2016): 188-192.

[33] Singhal, Nitesh, Parijat Sinha, Nitin Agarwal, and Muthukumar Murugesan. "Systems and methods for facitiating card verification over a network." U.S. Patent Application 12/819,774, filed December 22, 2011.

[34] Murugesan, Muthukumar, and T. Ravichandran. "Evaluate database compression performance and parallel backup." International Journal of Database Management Systems 5.4 (2013): 17.

[35] Muthukumar, M., and T. Ravichandran. "Analyzing compression performance for real time database systems." Int. Conf. on Advanced Computer Engineering and Applications (ICACEA). 2012.

[36] Murugesan, Muthukumar, K. Karthikeyan, and K. Sivakumar. "Novel investigation methodologies to identify the SQL server query performance." Indian Journal of Science and Technology 8.27 (2015).

[37] Murugesan, C., and T. Ravichandran. "Real time database compression optimization using iterative length compression algorithm." Int. Conf. on Computer Science and Information Technology, USA. 2013.

[38] Muthukumar, M., and T. Ravichandran. "Optimizing multi storage parallel backup for real time database systems." IJESAT, ISSN: 2250-3676.

[39] Muthukumar, M., and T. Ravichandran. "Optimizing and enhancing parallel multi storage backup compression for real-time database systems." International Journal of Computer Technology and Applications 3.4 (2012).

[40] Murugesan, Muthukumar, and T. Ravichandran. "Performance Enhancement Evaluation in Database Decompression Using HIRAC Algorithm." International Journal of Computer Science Issues (IJCSI) 9.6 (2012): 35.

[41] M. MUTHUKUMAR, Dr.T. RAVICHANDRAN." Database Compression Performance Enrichment using HIRAC Algorithm", 2012, Karpagam University Research Congress - 2012 (KURC 2012).

[42] M. MUTHUKUMAR, Dr.T. RAVICHANDRAN," Enhanced Database Compression and Decompression Techniques for Performance Improvement", 2013, State Level Seminar on "EMERGING TRENDS AND ISSUES", Kongu Arts and Science Collegs. ERODE.

[43] Dr.Muthukumar Murugesan, Dr. K. Karthikeyan, Dr.K. Sivakumar, "Analyzing Integral Components of SQL Server Databases", International Journal of Applied Engineering Research (IJAER), Volume: 10, Issue.9, Page(s): 24189-24200, 2015.

[44] K.Karthikeyan M Muthukumar, Senthil Pandian, "Analyzing and Improving the Performance of Decision Database with Enhanced Momentous Data Types", Asia Journal of Information Technology, Volume: 16, Issue.9, Page(s): 699-705, 2017.