IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Analysis of Epidemic Diseases Using Big Data Analytics

Y. Deepthi[1*], A. Radhika[2], Ch. Praneeth[3]

[1]Computer Science and Engineering, SRK Institute of Technology, JNTU Kakinada University, Vijayawada, India
[2]Computer Science and Engineering, SRK Institute of Technology, JNTU Kakinada University, Vijayawada, India
[3]Computer Science and Engineering, SRK Institute of Technology, JNTU Kakinada University, Vijayawada, India

Available online at: www.ijcseonline.org

Accepted: 12/Jul/2018, Published: 31/Jul/2018

***Abstract*—** There are a number of epidemic diseases such as Ebola Virus, Zika Virus, Dengue, Malaria etc that are spreading all over the World. It is necessary to provide awareness about these contiguous diseases to the people. To provide this, a thorough analysis is done on all these diseases and analysis is done on the type of people who effected mostly due to certain climatic conditions and country they are living in. For epidemic diseases analysis, R programming plays a vital role in data science and analysis. Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve complex problems analytically. The "Analysis of Epidemic Diseases" is an application which provides an opportunity for various Countries to estimate the severity of occurrence of various diseases, death counts etc in the coming years on the basis of previous Countries statistics, climatic conditions, death rates, confirmed or suspected cases and so on.

***Keywords*—** Epidemic, Zika, Dengue, Malaria, Ebola Virus, Analysis, Statistics, Prediction

I. INTRODUCTION

Till now the statistics of various diseases in different Countries are being done manually using paper and pen or by using Microsoft Excel. There are a number of epidemic diseases such as Ebola Virus, Zika Virus, Dengue, Malaria etc that are spreading all over the World. It is necessary to provide awareness about these contiguous diseases to the people. To provide this, a thorough analysis is done on all these diseases and analysis is done on the type of people who effected mostly due to certain climatic conditions and country they are living in.

For epidemic diseases analysis, R programming plays a vital role in data science and analysis. Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve complex problems analytically. The "Analysis of Epidemic Diseases" is an application which provides an opportunity for various Countries to estimate the severity of the disease occurrence, death counts etc in the coming years on the basis of previous Countries statistics, climatic conditions, death rates, confirmed or suspected cases and so on. The following diseases were explained briefly as follows.

Ebola:

Ebola virus disease (EVD), formerly known as Ebola hemorrhagic fever, is a severe, often fatal illness in humans. The virus is transmitted to people from wild animals and spreads in the human population through human-to-human transmission. The average EVD case fatality rate is around 50%. Ebola virus disease (EVD) first appeared in 1976 in South Sudan and the Democratic Republic of Congo. The latter occurred in a village near the Ebola River, from which the disease takes its name.

Dengue:

Dengue is a mosquito-borne viral infection. The infection causes flu-like illness and occasionally develops into a potentially lethal complication called severe dengue. The global incidence of dengue has grown dramatically in recent decades. About half of the world's population is now at risk. Dengue is found in tropical and sub-tropical climates worldwide, mostly in urban and semi-urban areas.

Zika Virus:

Zika Virus is a mosquito-borne flavivirus that was first identified in Uganda in 1947 in monkeys through a network that monitored yellow fever. It was later identified in humans in 1952 in Uganda and the United Republic of Tanzania. Zika Virus disease is usually mild and requires no specific treatment. If symptoms worsen, they should seek medical care and advice. There is currently no vaccine available.

Malaria:

Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female Anopheles mosquitoes called "malaria vectors". There are 5 parasite species that cause malaria in humans, and 2 of these species – P. falciparum and P. vivax – pose the greatest

threat. P. falciparum is the most prevalent malaria parasite on the African continent. It is responsible for most malaria-related deaths globally.

As the above-mentioned diseases are epidemic i.e., more infectious and spread rapidly among individuals in an area or population at the same time, which causes a rigorous effect all over the world. Hence, it is necessary to analyze these epidemic diseases. With the help of these analyses, we can do some kind of favor to the society. As the analysis provides information regarding the extent of severity, its effects, and consequences, the following measures can be taken for reducing the level of infections all over the world.

We present the remainder of this paper as follows. We describe the technology which was used in our analysis in Section II. In Section III, the system model and problem statement are introduced and the new protocol for analysis and forecasting is proposed as well. Then, we implement the analysis and forecasting using R tool by considering various datasets and experiment results are presented in Section IV, followed by the conclusion in Section V.

II. TECHNOLOGY DESCRIPTION

A. *R Programming Language*

R is a programming language and software environment for statistical analysis, graphics representation, and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team [3]. The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. R is free software distributed under a GNU-style copyleft, and an official part of the GNU project called GNU S.

B. *Evolution of R*

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. R made its first appearance in 1993.

- A large group of individuals has contributed to R by sending code and bug reports.
- Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive.

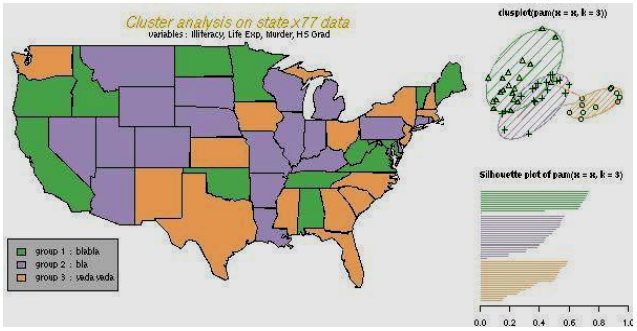

Figure 1. Graphical Representation in R Studio

C. *Features of R.*

As stated earlier, R is a programming language and software environment for statistical analysis, graphics representation, and reporting. The following are the important features of R:

- R is a well-developed, simple and effective programming language R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors, and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing on the papers.

As a conclusion, R is world's most widely used statistics programming language. It's the # 1 choice of data scientists and supported by a vibrant and talented community of contributors.

D. *RStudio*

RStudio allows the user to run R in a more user-friendly environment. It is open source (i.e. free).

III. SYSTEM MODEL

A. *Existing System*

In the existing system, we found the data regarding these diseases in an Excel format, which won't give a clear analysis to the people. We cannot apply different types of test cases and analysis on the Excel-based data. Through this system, we can make the analysis on simple data but it becomes hectic when there is a need for making analysis, forecasting etc on the larger datasets.

Disadvantages

- It requires more manpower to complete the task.
- Space & Time complexities will be far more for the huge amount of data.
- Vulnerable To Fraud

- May lead to human errors while performing analysis
- May lead to incorrect results
- Difficult to analyze / test
- Incapable of Supporting Quick Decision Making
- Not Designed for Collaborative Work (planning, forecasting, analyzing)

B. Proposed System

The system i.e., "Analysis of Epidemic Diseases" is used to overcome the drawbacks of an existing system. As we cannot provide a clear analysis through XML Spreadsheets, we designed a proposed system. In order to give a clear picture about these epidemic diseases, we are going to implement analysis such as the type of people who are affected mostly under certain climatic conditions or which country is highly affected over the period of time and so on using the R-programming tool [4]. By using this method we can visualize our data analysis and predictions using different types of diagrammatic formats such as graphs, pie-charts and so on which makes our analysis easier. It will provide the way to analyze the epidemic diseases data and represents the data in graphical Representation.

Advantages

- R is open source
- R supports extensions and can run anywhere
- R provides various packages for data analysis
- Space & Time complexities will be less when compared to other programming languages
- Minimum number of lines of code
- Easy to understand the data
- Visualization is possible with R
- Outstanding graphical outputs
- Future predictions based on data can be done in R

IV. IMPLEMENTATION

A. Modules Description

1) Data Preparation

In the first step, we have to set up an Analytic Sandbox i.e., Workspace. Familiarize with data sources is the more important task, that means, should have a clear picture of what data is needed and what data is available for analysis. Now perform ELT---- extract the data and determine the required transformations, data connections for raw data. Consider the statistically useful measures and check for data quality. Finally, perform Data Cleaning Process (keep data which is needed for the analysis). We are using inbuilt "tm" package available in R for Data Cleaning (Text Mining). Several methods are available for removing numbers, capitalization, common words, punctuation, whitespaces, stop words (a, and, also, are) and prepare the dataset for analysis.

- tm_map () is used for applying transformations to the Corpus data. This is available in "tm" package respectively [2].
- gsub() method is used to remove unwanted words or expressions in datasets.

Stemming is another part in Data Cleaning. It removes the duplicate words from the data.

2) Model Planning

In this module, we have to determine the methods and techniques based on hypothesis, data structure etc to meet the requirements. Data exploration is the first step in data analysis which involves summarization of main characteristics of a dataset. Here we explore the data using more advanced statistical software i.e., R. Explore the relation of the data to the underlying subject. This exploration involves various steps such as convert a variable to different data type, transpose a Dataset, sorting, creating plots, finding frequencies and so on. Some methods are given below as follows—

- table() ---- gives frequencies of particular data
- unique() ---- removes duplicate
- is.na() ---- helps to identify missing values
- y[is.na(y)] <- mean(y,na.rm=TRUE) --- replaces Null values (NA) with the mean value.

Variable selection is done through Feature Selection method. The basic idea is to keep most relevant but not redundant feature for a predictive model that can yield optimal accuracy.

Create a model to summarize understanding of how the data relates to the end goal. ANOVA() [3] is one of the inbuilt models in R. It refers to the analysis of variance (ANOVA) and is a statistical procedure used to test the degree to which two or more groups vary or differ in an experiment. Finally the selection of a model for prediction.

We will consider Time Series Modeling for prediction and forecasting. It involves working on time (years, month, hour, minutes) based data. It helps to derive hidden insights to make informed decision making. It is mainly used to identify a trend or detrend of the data structure. The inbuilt "ts" package is used for creating time series object. In order to predict we need a training dataset, which is used to discover potentially predictive relationships. Then the time series object is sent to ARIMA() function where it can be used to fit an autoregressive integrated moving averages (imperfectly predictable) time-series model to past values of a time series. ARIMA() is a generalization of ARMA(). Now forecasting is done with the help of "forecast" package in "R". The ets() function in the forecast package, can be used to fit exponential models.

3) Model Building

In this phase, we develop the datasets which are sufficiently robust for the model and analytical techniques. "rpart" package is used for model building in R. The training set is considered for initial experiments. Validating using smaller datasets is done. For validating we use the inbuilt "Bayesian classifier" technique respectively. The "e1071 package" [3] contains a function named "naiveBayes()", which is helpful in performing Bayes classification. The function is able to receive categorical data and contingency table as input and returns an object of class "naiveBayes". This object can be passed to "predict()" to predict outcomes of unlabeled subjects.

4) Communicate Results

Interpret the results by using the available data. Finally, we summarize the findings / outputs based on the conditions. Visualize the data using various methods in R tool. For effective outputs, we use the "ggplot2" package for visualizing the data. Here "qplot()" method is used.

- qplot() ----- This function can be used to create the most common graph types. While it does not expose ggplot's full power, it can create a very wide range of useful plots.

B. Datasets

ebola.R × ebol ×

Filter

	Country	Year	Month	Climate	Male	Female	A14	A44	A45
1	Guinea	2014	Dec-17	Hot	1017	1090	329	1216	545
2	Guinea	2014	Dec-24	Hot	1156	1218	371	1360	622
3	Guinea	2014	Dec-31	Hot	1156	1218	371	1360	622
4	Guinea	2015	Jan-07	Hot	1309	1410	431	1539	727
5	Guinea	2015	Jan-14	Hot	1334	1430	442	1564	736
6	Guinea	2015	Jan-21	Hot	1341	1438	443	1572	742
7	Guinea	2015	Jan-28	Hot	1341	1438	443	1572	742
8	Guinea	2015	Feb-04	Hot	1413	1508	460	1648	791
9	Guinea	2015	Feb-11	Hot	1413	1508	460	1648	791
10	Guinea	2015	Feb-18	Hot	1481	1586	487	1720	837
11	Guinea	2015	Feb-25	Hot	1481	1586	487	1720	837
12	Guinea	2015	Mar-04	Hot	1532	1637	495	1776	878

Figure 2. Ebola Dataset

malariadata.R × malaria × new × lz × tz ×

Filter

	Country	Year	Climate	Type.of.Case	Deaths
1	Afghanistan	2014	Arid and Semiarid	Malaria - number of reported deaths	32
2	Algeria	2014	Hot	Malaria - number of reported deaths	0
3	Angola	2014	Rainy and Dry	Malaria - number of reported deaths	5 714
4	Argentina	2014	Hot and Humid	Malaria - number of reported deaths	0
5	Armenia	2014	Dry	Malaria - number of reported deaths	
6	Azerbaijan	2014	Hot and Cold	Malaria - number of reported deaths	0
7	Bahamas	2014	Hot	Malaria - number of reported deaths	
8	Bangladesh	2014	Subtropical	Malaria - number of reported deaths	45
9	Belize	2014	Subtropical	Malaria - number of reported deaths	0
10	Benin	2014	Hot and Humid	Malaria - number of reported deaths	1 869
11	Bhutan	2014	Hot and Humid	Malaria - number of reported deaths	0
12	Bolivia	2014	Tropical	Malaria - number of reported deaths	1

Figure 3. Malaria Dataset

dengue.R × dengue × lz × tz × new × new × dengue timeseries.R ×

Filter

	Region	Country	Year	climate	Population	Deaths
1	Andean Subregion	Bolivia	2017	tropical	1,10,52,864	2
2	Andean Subregion	Colombia	2017	tropical	4,90,67,981	47
3	Andean Subregion	Ecuador	2017	temperate	1,66,25,776	1
4	Andean Subregion	Peru	2017	tropical	3,21,66,473	81
5	Andean Subregion	Venezuela	2017	tropical marine	3,19,25,705	13
6	Central America Ithsmus and Mexico	Belize	2017	sub tropical	3,74,651	0
7	Central America Ithsmus and Mexico	Costa Rica	2017	sub tropical	49,05,626	0
8	Central America Ithsmus and Mexico	El Salvador	2017	humid tropical	61,67,147	0
9	Central America Ithsmus and Mexico	Guatemala	2017	warm and humid	1,70,05,497	14
10	Central America Ithsmus and Mexico	Honduras	2017	hot and humid	83,04,677	0
11	Central America Ithsmus and Mexico	Mexico	2017	sub-arctic and arctic	13,02,22,815	32

Figure 4. Dengue Dataset

zika datacode.R × zika ×

Filter

	Year	Month	Date	Climate	Country	Suspected	Confirmed	Importedcases
56	2016	September	8	hot humid	Mexico	0	2,388	15
57	2016	September	8	sub tropical	Belize	0	5	0
58	2016	September	8	hot and dry	Costa Rica	1,721	946	32
59	2016	September	8	tropical	El Salvador	11,098	51	0
60	2016	September	8	tropical	Guatemala	2,473	442	0
61	2016	September	8	tropical	Honduras	30,735	225	0
62	2016	September	8	tropical	Nicaragua	0	1,755	3
63	2016	September	8	tropical maritime	Panama	1,508	294	40
64	2016	September	8	sub tropical	Cuba	0	3	30
65	2016	September	8	tropical	Dominican Republic	5,109	318	0
66	2016	September	8	tropical	French Guiana	9,565	483	10

Figure 5. Zika Virus Dataset

C. Results

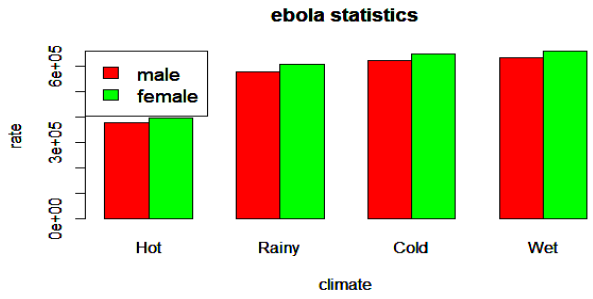

Figure 6. Ebola Analysis

From Figure.6, we analyzed the death count of Male as well as Female based on Climate in different Countries. By this, we concluded that the death rate is high in wet climatic conditions for Ebola.

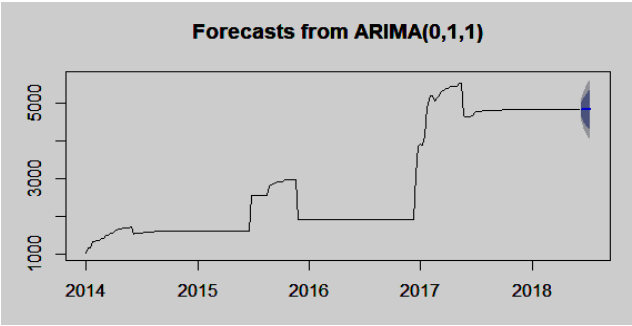

Figure 7. Ebola Time Series

From Figure.7, we can observe the predictions of Ebola disease in various countries in future. It is clear that the death rate is going to decrease in future when compared with the 2017 statistics for Ebola.

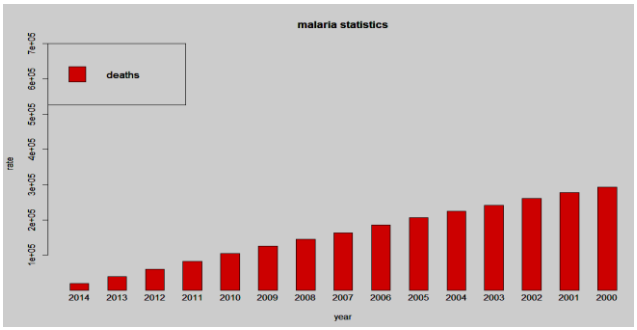

Figure 8. Malaria Analysis

From Figure.8, we can observe that the death rate of Malaria is decreasing gradually when one observes from the year 2000 to 2014 respectively.

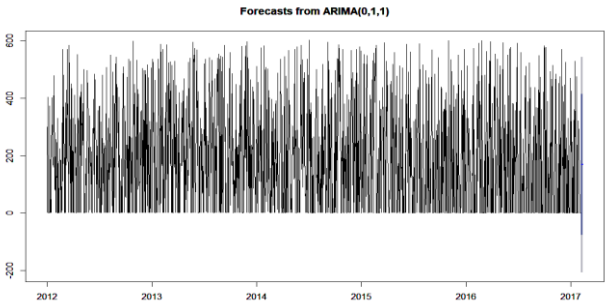

Figure 9. Malaria Time Series

From Figure.9, we can observe the predictions of Malaria disease in various countries in future. It is clear that the death rate is going to decrease in future when compared with the previous statistics from the year 2012 to 2017 for Malaria.

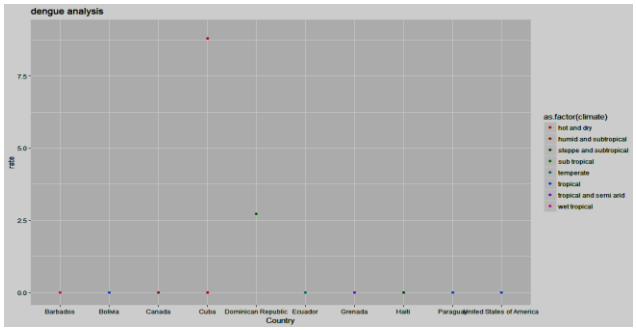

Figure 10. Dengue Analysis

From Figure.10, we calculated the death rate by considering Climate and Country. So from the graph, it is clear that Cuba has the high death rate in hot and dry climate. Hence Cuba is highly affected Country based on our analysis. Hence Cuba is more vulnerable to Dengue.

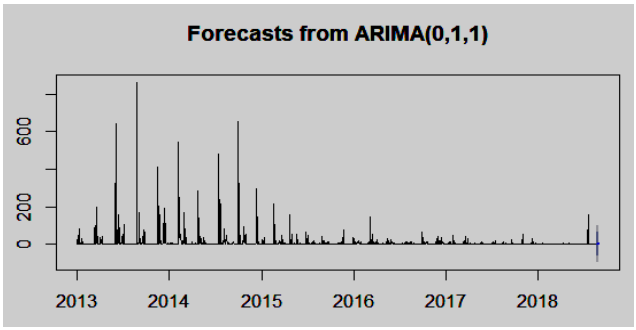

Figure 11. Dengue Time Series

From Figure.11, we can observe the predictions of Dengue disease in various countries in future. It is clear that the death rate is going to be decreased in future for Dengue. We can also say that the effect of this disease is very minimal in 2017 when compared to the previous years.

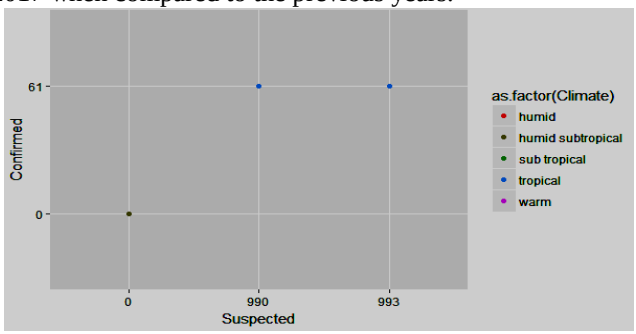

Figure 12. Zika Virus Analysis

From Figure.12, we calculated the death rate by considering Confirmed and Suspected. So from the graph, it is clear that there is a high number of suspected and confirmed cases in the tropical climate.

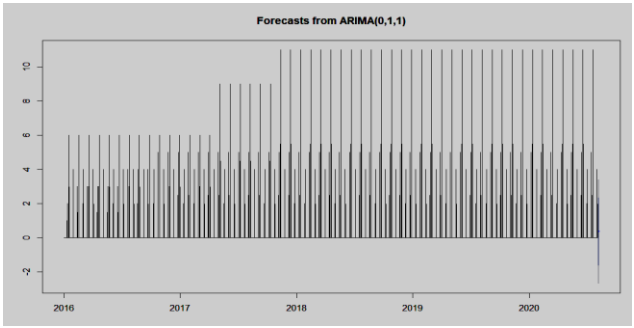

Figure 13. Zika Virus Time Series

From Figure.13, we can observe the predictions of Zika Virus disease in various countries in future. It is clear that the death rate is going to decrease in future. We can also say that the effect of this disease is very minimal in the coming years, say, 2019, 2020, when compared to the previous years.

V. CONCLUSION

The Analysis of Epidemic disease application is developed keeping in view to reduce the death rate caused due to epidemic diseases all over the World. It is useful and user-friendly for the people, as it provides the information related to the effects of epidemic diseases in different regions and also helps the Governments of various Countries to take the respective preventive measures based on the analysis and future predictions of these diseases. Hence death rate will be minimized. It also provides the graphical representation of the analyzed data.

REFERENCES

[1] Pressman RS, "Software Engineering", McGraw Hill Education, India, pp. 366-398, 2000.
[2] Richard E Fairley, "Software engineering concepts", McGraw Hill Education, India, pp. 456-520, 2001.
[3] Richard Cotton, "Learning R: A Step-by-Step Function Guide to Data Analysis", O Reilly Media, USA, pp.250-300, 2013.
[4] Jared P. Lander, "R for Everyone: Advanced Analytics and Graphics", Addison-Wesley Data and Analytics, USA, pp. 300-385, 2017.
[5] R. K. Bathla, Jitender Nath Srivastva, "An Ethical Approach of Big Data & Machine Learning Using Innovation of Python", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.8, Issue.6, pp.1-9, 2018.
[6] P. Meenakshi, M. Veeresh Babu, "Load Prediction for Resource Management in Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering, pp.340-346, Vol.6, Issue.8, pp.1-9, 2016.

Authors Profile

Ms. Y. Deepthi, pursed Bachelor of Technology from NRI Institute of Technology affiliated to JNTUK University in the year 2016. She is currently pursuing Master of Technology from SRK Institute of Technology, India.

Mrs. A. Radhika, working as Senior Assistant Professor in SRK Institute of Technology. She published 12 papers in different International journals and attended nearly 7 conferences. Her area of interest is Computer Network, Information security, R programming.

Mr. Ch. Praneeth, working as Assistant Professor in SRK Institute of Technology. He published 4 papers in different International Journals. His area of interest includes Big Data Analytics, Cloud computing.

