ML Based HCC Survival Prediction System

Jayaraman Vikas, G Vijendar Reddy, N V Ganapathi Raju, A Sai Hanuman, Lakshmi Sushma Kolli

Abstract: ***Data Science is an interdisciplinary branch of technology, which deals with extraction of knowledge and insights from large amounts of data. It combines different fields of work of statistics and computation for data interpretation to facilitate decision making. Hepatocellular Carcinoma (HCC) is any individual who is hepatitis C positive. It is most common type of primary cancer in adults. If observed from 1980, the incidence of HCC is almost tripled. In 2018, around 42,220 adults have been estimated to have been dead. Among the affected, it was observed that more men were affected with HCC than women. This was more common in African and Asian continents. The death of a person disturbs the stability at home. This paper proposes to use data science and machine learning to build a system that may not help in predicting how long one will survive, but to find how much the treatment can be successful for him/her. Also, it provides an insight on how many people are alive with the same stage of disease and also throw light on the effectiveness of the treatment methodologies. This project performs a comparative study on various ML classifiers to identify the best one and has found Logistic Regression to offer the best performance with an accuracy of 99.49%.***

Keywords : HCC, Machine Learning, Python, Survival Prediction

I. INTRODUCTION

From an era of doing every task by hand to an era of automating everything, the humans have had a wonderful transition. It all started with a vision to reduce human effort and make their life easier and today its reach and horizons are boundless. Machines have always been known to create a revolution. The world witnessed the Industrial Revolution in 18th Century for the first time when all the tough labor of human was replaced with specialized machines for the job. No one could ever imagine, then, that machines could not just reduce effort and act according to our orders but also start acting on their own without our intervention. ML, though in existence from a long-time, has started receiving its due lately. People have realized the power of ML and it is slowly becoming a ubiquitous part of our lives.

Revised Manuscript Received on April 06, 2019.

Jayaraman Vikas, Bachelor of Technology in Information Technology from Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

Mr. G Vijendar Reddy,Associate Professor of Information Technology, pursing Ph.D in Computer Science Engineering from JNTU, Hyderabad, India.

Dr. N V Ganapathi Raju working as Professor, in the department of IT, GRIET. Completed Ph.D. in CSE from Jawaharlal Nehru Technological University Kakinada, India

Dr. Akundi Sai Hanuman, Professor of Computer Science and Engineering, completed his Ph.D. from Acharya Nagarjuna University, India

Ms. Lakshmi Sushma Kolli, Assistant Professor of Information Technology, completed her M.Tech from KL University, India

Doesn't every individual love to be cared and pampered to his needs regardless of asking for it. Humans have been created with an ability to learn and evolve with experience without being explicitly trained. This thought struck the mind of some person back then and led to a proposal of machines' learning from our actions. Unlike humans who learn with knowledge as a base, Machines are made to learn using a data-driven approach. Histological data of our actions act as input for machines to learn and develop trends. These trends enable them to make decisions and act on their own. Machine learning is a branch of AI. AI aimed to build more intelligent systems but could not go beyond finding shortest path between two points and other basic tasks. It got highly difficult to code and solve complex problems and evolving challenges which led to the realization of the machine learning by itself. In the year 1952, Arthur Samuel devised the first computer learning program - A Checkers game. The IBM machine got better after every game it played. It traced the moves that were best winning strategies and included them to its program. In 1959, he was the first person to coin the term – "Machine Learning". Now, ML has become ubiquitous and is spread in various walks of technology. On this planet, it is only possible for humans to find patterns in data. But as the volume of data goes up, it gets difficult and time taking for humans to compute. This is when ML comes to the rescue, helping people with large data in minimum amount of time. In 1967, Nearest Neighbors algorithm was introduced which pioneered the basic pattern recognition capabilities in machines. In 1981, Gerald Dejong proposed the concept of Explanation Based Learning (EBL) for the first time. By this the computers could analyze training data and create a general rule which it follows to discard irrelevant data. In the 90s, ML slowly started drifting from knowledge driven approach to data driven approach. Scientists created computer programs that enabled computers to draw conclusions or learn from data. Right from the Personal Assistants in our phones to online frauds, to search engine result refining, to video surveillance, to recommendation systems, to spam filtering, there are several use cases in which ML is playing a vital role. Among all these, one important use case of ML is healthcare. Machine Learning is helping to detect various diseases and help offer assistance insight onto effectiveness of the treatment methodologies. ML is helping to ease the job in diagnoses of many diseases which earlier needed microscopes and other equipment on the part of pathologists. Google has developed its own ML algorithm to identify cancerous tumors on mammograms. Stanford also proposed a Deep learning algorithm for the detection of skin cancer. With advancements in ML, one can expect more effective and reliable applications to pave its way in diagnostics and clinical decision making.

With all this, it can be definitely said that ML has put another arrow in the quiver of medical diagnoses.

II. RELATED STUDY

Hepatocellular Carcinoma (HCC) is one of the most common primary liver cancers in the world that originate in the liver and spread to other parts of the body. It is different from secondary liver cancers that originate at some other organ and spread to the liver. Populations in East Asia and Pacific, South Asia, and parts of Sub-Saharan Africa are more susceptible to HCC, largely due to the outbreak of infection decades ago. Like many other cancers, if detected at an early stage, it can be cured. In case of advanced stage of disease, it cannot be cured but medications can help to support and prolong the life. Annually, around 0.7 million people contract HCC and 0.6 million succumb to the disease.

Although the doctors are not very sure about the causes of HCC, they have identified a few factors that can catalyze the chances of getting affected with it. The most important and major factors influencing HCC occurrence are the Hepatitis B or Hepatitis C infections. After years of being infected with these viruses, one may end up contracting HCC. Hepatitis B virus is usually self-cured, i.e. 95% of people get rid of it after a few weeks. The remaining 5% become carriers and are responsible for spreading the disease. Obesity is another factor that escalates the chances of contracting HCC. It can lead to causing Non-Alcoholic Fatty Liver Disease (NAFLD) which quadruples the chances of occurrence of HCC. Diabetes leads to high insulin levels in the body and causes damage to the liver. Hence, it also increases the chances of one getting affected with HCC by three folds. Cirrhosis – a serious disease in which liver cells get damaged and replaced by scar tissues. Hepatitis B or Hepatitis C infection, alcohol consumption, certain drugs and iron storage disease are factors for occurrence of Cirrhosis. Annually 2.5% of the cirrhosis patients will contract HCC. Alcohol consumption and smoking are risk factors for occurrence of most liver diseases and HCC is no exception. It is estimated that alcohol consumption greater or equal to 80g/day increases the risk of HCC by 75% and between 40g to 79g per day increases it by 37%. But few researches suggest that social or occasional drinking and smoking is as risky as regular drinking and smoking. Hemochromatosis AKA Iron storage disease is the excessive storage of iron in the liver which causes damage and leads to contraction of HCC. The risk of HCC is around 20 times higher in patients with hemochromatosis. Aflatoxin, which is abundantly found in corn-based diet, is another important catalyst for HCC occurrence. The U.S. has safety measures that control the amount of aflatoxin in food supply. China moved from corn-based to rice-based diet due to increased mortality rates associated with risks of aflatoxin in diet. Gender is another determinant of HCC, with more men being affected than women. It was also observed that the GDP of a nation helped in reducing the occurrences of diseases, as it helped the nations spend money on medical researches and health schemes. It was observed that all these risk factors caused as many non-cancer deaths as cancer deaths. Hence prevention of these would help in reducing the HCC occurrence and also other medical complications.

In the early stages, one may not see any symptoms but one may notice one or more of the following symptoms as the cancer progresses – pain in the upper right part of belly, a lump in the upper belly, loss of appetite, weight loss, fatigue, nausea, yellow skin and eyes, pale chalky bowel and dark urine and fever.Doctors might use blood tests, imaging tests, liver biopsy and other tests to diagnose HCC.

Various treatments are available for HCC. Radiation is one form of treatment and is of two variants. Internal radiation involves blocking flow of blood to the artery carrying blood to the tumor and thus killing those cancerous cells. External radiation involves targeting beams of radiation onto the target spot and burning the cancerous cells. Chemotherapy is another form of treatment. The only issue of concern is the side effects like nausea, fatigue, bruising which might go away over the course of treatment and proper medication. Alcohol injection is another course of treatment where ethanol is injected into the targeted cells to burn them off. This procedure is performed under local anesthesia. Cryoablation and Radiofrequency ablation involve two methods which kill the cancerous cells by freezing or heating the cells with a thin metal probe. Liver transplant and removing part of liver are also forms of treatment.

III. PROPOSED SYSTEM

Following the conventional approach to get the best results, we started our job with data collection. Any ML based system is worthless without real-time histological data. But at the same time, it is a well-known fact that creation of data is a time-taking and expensive affair. So, the most optimal approach of using a real-time data from online repositories (UCI Repository) was adopted for this research.

Next, the features that least influenced the prediction process had to be identified and eliminated. For this, without blindly following the technical approach, the insight gained from the literary survey was put to use to conclude that though there were only a few major factors for prediction, yet all the 49 features had some profound influence clinically in a few cases which made them vital to the prediction process.

The further step included cleaning the data. The dataset collected had missing values which needed to be cleaned before analysis. Missing data fields can occur due to various reasons and can be handled by various approaches. The data collected for this project had 26 categorical features and 23 continuous features. This system uses Standard Imputer to impute the categorical features with the mode of that feature and the continuous features are imputed using mean value of that feature.

After cleaning the data, the next job in hand was to transform the data. Each feature in the dataset had a different range. If not scaled, the features with higher range would bias the prediction process thus impacting the accuracy of the job-in-hand. Hence, Standard Scaler has been applied to scale all the features.

The final step was to apply various algorithms and perform a comparative study to identify the best and the most accurate algorithm. Six different classification algorithms – Logistic Regression, Gaussian Naïve Bayes, Support Vector Classifier, Stochastic Gradient Classifier, K Neighbors and Random Forest, were included in the comparative study and Logistic Regression is found to offer the best performance.

The key components of the system are:

a) **Imputer:** The first core challenge with the dataset was the missing values. To overcome this issue,

the imputer imputes the missing values using mean or mode based on the nature of the feature. Continuous features are imputed using mean value of the feature and whereas coming to Nominal or Categorical features, they are imputed using mode.

b) **Scaler:** The ranges of data of the different features vary greatly. This may bias the output of prediction and hence needs to be handled. The Scaler normalizes the features and transforms all of them to a common range.

c) Classifier: The classifier is responsible for applying the Classification algorithm on the data and predicting the survival rate of patient for one year.

The below figure (Fig. 1), represents the flow of the Survival prediction system.

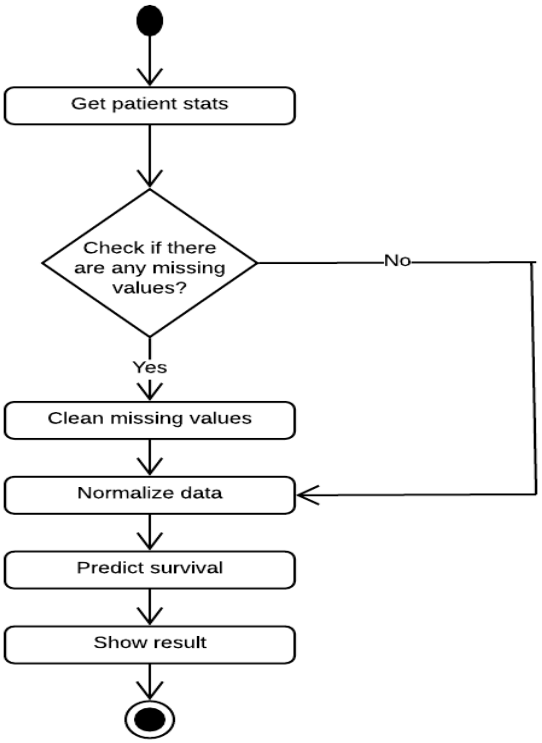

Figure 1. Flow of proposed system

IV. RESULT & ANALYSIS

The below figure (Fig.2) represents a snapshot of the dataset that has been used in this project.

	Gender	Symptoms	Alcohol	HBSA	HBeA	HBCA	HCVA	Cirrhosis	EC	Smoking	...	AP	TP	Creatinine	NumNod	MdN	DB	Iron	OS	Ferritin
0	1	0.0	1	0.0	0.0	0.0	0.0	1	0.0	1.0	...	150.0	7.1	0.70	1.0	3.5	0.5	NaN	NaN	NaN
1	0	NaN	0	0.0	0.0	0.0	1.0	1	NaN	NaN	...	NaN	NaN	NaN	1.0	1.8	NaN	NaN	NaN	NaN
2	1	0.0	1	1.0	0.0	1.0	0.0	1	0.0	1.0	...	109.0	7.0	2.10	5.0	13.0	0.1	28.0	6.0	16.0
3	1	1.0	1	0.0	0.0	0.0	0.0	1	0.0	1.0	...	174.0	8.1	1.11	2.0	15.7	0.2	NaN	NaN	NaN
4	1	1.0	1	1.0	0.0	1.0	0.0	1	0.0	1.0	...	109.0	6.9	1.80	1.0	9.0	NaN	59.0	15.0	22.0

Figure 2. Snippet of Data

The below figure (Fig.3) represents the correlation heat-map.

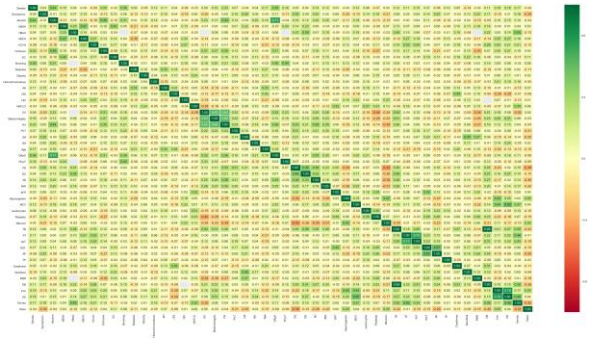

Figure 3. Correlation Heat Map

The below figure (Fig. 4 to Fig. 8) represent histograms that were used to understand trends from the dataset.

Fig. 4 plots the relation between Age group and incidence of HCC.

Fig. 5 plots the relation between Alcohol consumption everyday (in grams) and incidence of HCC.

Fig. 6 plots the relation between Packs of cigarette consumed annually and incidence of HCC.

Fig. 7 plots the relation between Gender and HCC mortality.

Fig. 8 plots the relation between HCC incidence and HCC mortality.

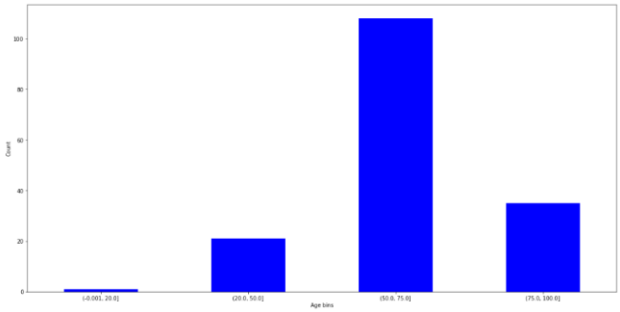

Figure 4. Age bins Vs HCC Incidence

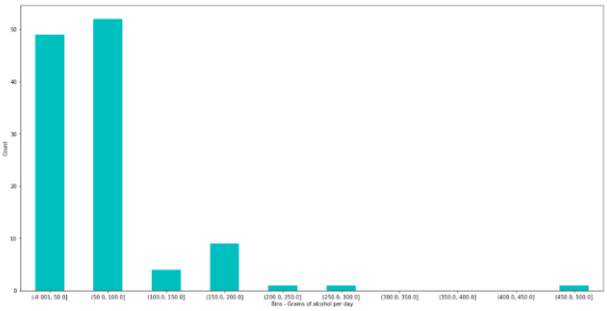

Figure 5. Grams of Alcohol per day Vs HCC Incidence

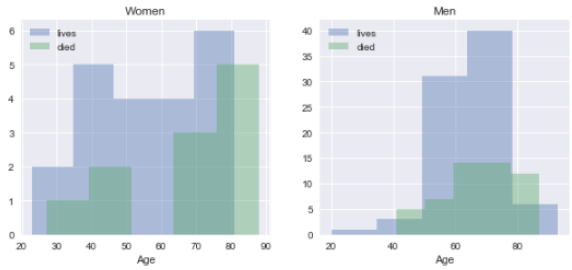

Figure 6. Gender Vs HCC Mortality

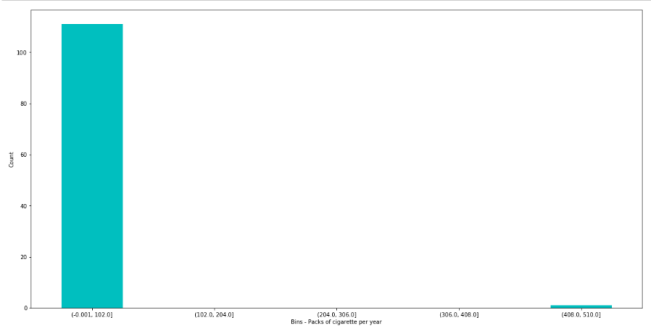

Figure 7. Packs of Cigarette per year Vs HCC Incidence

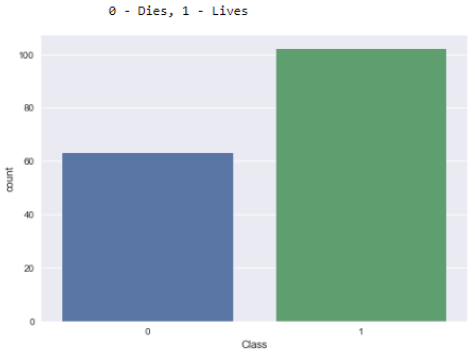

Figure 8. HCC Mortality Vs HCC Incidence

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        63
           1       1.00      0.99      1.00       102

avg / total       0.99      0.99      0.99       165

LogisticRegression: F1 after 5-fold cross-validation: 99.49% (+/- 0.02%)
```

Figure 9. Logistic Regression Classifier

```
Cross-validated scores: [0.93333333 1.         1.         1.         1.        ]

              precision    recall  f1-score   support

           0       1.00      0.95      0.98        63
           1       0.97      1.00      0.99       102

avg / total       0.98      0.98      0.98       165

Gaussian Naive Bayes: F1 after 5-fold cross-validation: 98.67% (+/- 0.05%)
```

Figure 10. Naive Bayes Classifier

```
Cross-validated scores: [0.97674419 0.95238095 0.97435897 0.97435897 0.97435897]

              precision    recall  f1-score   support

           0       0.94      0.97      0.95        63
           1       0.98      0.96      0.97       102

avg / total       0.96      0.96      0.96       165

SVC: F1 after 5-fold cross-validation: 97.04% (+/- 0.02%)
```

Figure 11. Support Vector Classifier

```
Cross-validated scores: [0.95238095 0.97560976 1.         0.97435897 1.        ]

              precision    recall  f1-score   support

           0       0.97      0.97      0.97        63
           1       0.98      0.98      0.98       102

avg / total       0.98      0.98      0.98       165

SGD: F1 after 5-fold cross-validation: 98.05% (+/- 0.04%)
```

Figure 12. Stochastic Gradient Classifier

```
Cross-validated scores: [0.78431373 0.81818182 0.85106383 0.82608696 0.85106383]

              precision    recall  f1-score   support

           0       0.84      0.43      0.57        63
           1       0.73      0.95      0.83       102

avg / total       0.77      0.75      0.73       165

KNN: F1 after 5-fold cross-validation: 82.61% (+/- 0.05%)
```

Figure 13. K Neighbours Classifier

```
Cross-validated scores: [0.92682927 0.95238095 0.95       0.97435897 1.        ]

              precision    recall  f1-score   support

           0       0.94      0.92      0.93        63
           1       0.95      0.96      0.96       102

avg / total       0.95      0.95      0.95       165

Random Forest: F1 after 5-fold cross-validation: 96.07% (+/- 0.05%)
```

Figure 14. Random Forest Classifier

The above figures (Fig.9 to Fig.14) denote the performance of various algorithms that were used in the comparative study. It is evident that Logistic Regression offered the best performance in terms of accuracy.

V. CONCLUSION

This paper has proposed and developed a system that can predict the survival of HCC patient using ML. It helps understand the effectiveness of treatment for the medical researchers. The dataset for this research has been taken from UCI repository. The dataset was collected at the Internal Medicine Service, Hospital and University Centre of Coimbra and contains one-year survival data of patients. It consists of 49 features. This paper has been able to achieve an accuracy level of 99.49% with 5 – fold cross validation using Standard Imputer, Standard Scaler and Logistic Regression Classifier. This system can be further enhanced to predict the amount of time patient can survive for and also suggest best course of treatment among the various treatments based on patients' stats.

REFERENCES

1. https://www.ncbi.nlm.nih.gov/books/NBK343640/
2. Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J Garcia-Laencina, Adelia Simao, Armando Carvalho, A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients, Journal of biomedical informatics, 58, 49-59, 2015.
3. https://www.webmd.com/cancer/hepatocellular-carcinoma
4. https://www.abc.net.au/news/health/2017-12-07/will-an-occasional-cigarette-damage-your-health/9087438
5. Data source:
6. https://archive.ics.uci.edu/ml/datasets/HCC+Survival
7. https://onlinelibrary.wiley.com/doi/full/10.1002/1097-0142(20000701)89:1%3C53::AID-CNCR8%3E3.0.CO%3B2-6
8. Takano S, Yokosuka Y, Imazeki F, Tagawa M, Omata M. Incidence of hepatocellular carcinoma in chronic hepatitis B and C: a prospective study of 251 patients. Hepatology 1995; 21; 650–5.
9. Ikeda K, Saitoh S, Koida I, Arase Y, Tsubota A, Chayama K, et al. A multivariate analysis of risk factors for hepatocellular carcinogenesis: a prospective observation of 795 patients with viral and alcoholic cirrhosis. Hepatology 1993; 18: 47–53.

10. Oka H, Kurioka N, Kim K, Kanno T, Kuroki T, Mizoguchi Y, et al. Prospective study of early detection of hepatocellular carcinoma in patients with cirrhosis. Hepatology 1990; 12: 680–7.
11. Chiba T, Matsuzaki Y, Abei M, Shoda J, Aikawa T, Tanaka N, et al. Multivariate analysis of risk factors for hepatocellular carcinoma in patients with hepatitis C virus related liver cirrhosis. J Gastroenterol 1996; 31: 552–8.
12. Haydon GH, Jarvis LM, Simmonds P, Harrison DJ, Garden OJ, Hayes PC. Association between chronic hepatitis C infection and hepatocellular carcinoma in a Scottish population. Gut 1997; 40: 128–32.
13. https://www.kaggle.com/miodeq/survival-prediction-with-logistic-regression-9

AUTHORS PROFILE

Jayaraman Vikas, is currently pursuing his final semester of Bachelor of Technology in Information Technology from Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India. His work Smart Lock Controlled using Phone call was published the previous year at ICSSIT, Tirunelveli and received the Best Paper Award.

Mr. G Vijendar Reddy,Associate Professor of Information Technology, pursing Ph.D in Computer Science Engineering from JNTU, Kakinada and M.Tech in Software Engineering from JNTU, Anantapur. I am having over 13 years of academic and research experience.

Dr. N V Ganapathi Raju working as Professor, in the department of IT, GRIET. Completed Ph.D. in CSE from Jawaharlal Nehru Technological University Kakinada and did Master of Technology in Computer Science and Technology from Andhra University, having 18 years of teaching experience which includes seven years of research experience in the area of Computer Science and Engineering. Research interests include Text Mining, Information Retrieval, NLP, Machine Learning and Data Science. The title of my research work is "Feature based Authorship Attribution". The result research work was published in various national and international journals including Scopus indexed, Free journals and Springer journals. Awarded a research funded project from University Grants Commission during 2014-16 and has been completed successfully.

Dr. Akundi Sai Hanuman, Professor of Computer Science and Engineering, completed his Ph.D. from Acharya Nagarjuna University, Guntur in 2012. He has over 22 years of experience in Academic, Industry and Research.
Dr. Akundi Sai Hanuman's Research interests include Data Clustering, Data Sciences, Machine Learning, Optimization Techniques and Distributed Systems.
Currently Dr. Sai Hanuman is acting as Dean of Examinations, GRIET since 2013. Previously he worked as Additional Controller of Examinations, Head of the Department, Chairman BOS and Convener for National Level cultural festival PULSE 2013.

Ms. Lakshmi Sushma Kolli, Assistant Professor of Information Technology, completed her M.Tech from KL University and has six years of academic experience. She had earned Bachelor of Technology of Engineering in Computer Science Engineering, and Masters of Technology of Engineering in Computer Science and Engineering. Ms. Sushma's research interests include data mining, opinion mining and cloud computing in which she has two publications, in various journals.