HEALTHCARE BIG DATA ANALYTICS USING BROWNBOOST CLASSIFIER BASED BLOOM HASH DATA STORAGE

R .VENKATESWARA REDDY [1], Dr. D MURALI [2], Dr. VINOD M VAZE [3]

Department of Computer Science and Engineering

1. Ph.D. Scholar, ShriJagdishprasadJhabarmalTibrewalaUniversity ,hunjhunu,Rajasthan-333010.

E-mail: venkatreddyvari@gmail.com.

2. Professor, CMR College Of engineering & Technology,Kandlakoya, Telangana, Hyderabad.

3. Professor, ShriJagdishprasadJhabarmalTibrewalaUniversity.Rajastan.

ABSTRACT

Healthcare big data analytics involves analysis of large amounts of patient data in order to discover useful information. There are many challenges that big data analytics presents in many areas, including cloud healthcare systems. Healthcare industry generates significant amounts of patient data. Recent research focuses on big data analytics-enabled models that increase prediction accuracy and reduce patient risk. Data storage is a concern. Data must be accessible from different locations within the distributed environment. We are studying a BrownBoost Classifier Based Bloom Hash Data Storage (BBC–BHDS) system to store and retrieve healthcare data from different locations in a distributed environment. This will allow for faster access and less space usage. Initial data collection (i.e. patient data) is done based on some parameters. The input data is then classified using the BrownBoost Classification (BBC). BrownBoost uses a non-convex probability loss function and the base SVM classifier to classify the patient data.

Keywords: BBC–BHDS, Big Data, Health care and SVM Classifier.

I. INTRODUCTION

Big data analytics has major challenges in several fields including cloud healthcare systems due to its significance. The healthcare industry produces large volume of patient data. Most of the recent research works aim at a big data analytics-enabled business model as it increases the prediction accuracy in the risk level of patients. However, the application of big data analytics in healthcare is still lagging due to the drawbacks in privacy of health information, security, budget constraints and handling a large volume of data with less complexity. The above issues are overcome by three different proposed techniques. Proposed in this research work have achieved enhanced data prediction accuracy with reduced false positive rate and complexity. Gramian symmetric data collection based random forest and regression classification technique was introduced for improving the disease data prediction accuracy in minimum time and less space complexities. For predicting patient data, a large volume of data was stored in the matrix form (i.e., rows and columns) with the utilization of Gramian Symmetric matrix. It helped minimizing the space complexity in big data predictive analysis. Next, random decision forest classifier was used for regression and classification. Here, the regression analysis was performed using bivariate correlation where the relationship between the variables was measured. Based on the correlation measure, multiple decision trees are constructed between the variables. Finally, all the decision trees were summed and the voting scheme was applied.Majority voting results was identified using the maximum argument function to attain accurate prediction in minimum time. This reduced false positive rate during the classification process in minimum time. Thus, the proposed GSDC-RFBRC technique improved prediction accuracy and minimizes the prediction time.

II. RESEARCH GAP

Future Health Condition Prediction (FHCP) algorithm was implemented for detecting the future health status of patients based on the prevailing health status with higher accuracy.Also, cloud-based MapReduce model was applied as the processing architecture for big data analysis.However, the data classification was not performed by using probabilistic data acquisition method.A data management methodology was discussed for optimizing the parallel execution of data-intensive bioinformatics workflows in a hybrid cloud.But, there was no efficiency seen in load balancing efficiency.The prediction accuracy of fuzzy model was increased but prediction accuracy with was not enhanced to the desired level.A new Convolutional Neural Network was designed on the basis of Multimodal Disease Risk Prediction (CNN-MDRP) algorithm with the aid of structured and unstructured data.A new Convolutional Neural Network was designed on the basis of Multimodal Disease Risk Prediction (CNN-MDRP) algorithm with the aid of structured and unstructured data.

III. PROBLEM ON HAND

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools like RDBMS, SQL Oracle etc. Big Data is a term given to large volumes of data that organizations store and process. However, it is becoming very difficult for organizations to store, retrieve and process the ever-increasing data. As healthcare industry generates large amounts of data continuously big data analytics plays a vital role in improving services and tackling some of the major challenges in health sector, especially regarding patient profile analysis, genomic analysis, public health monitoring, diagnosis and consulting, fraud analysis, Challenges with data collection and integrates with cloud servers. Data privacy and integrity over cloud server. Selection of Big Data Platforms and tools for Big Data Analytics over Real time healthcare data.

IV. Enhanced Gramian Symmetric Big Data Collection Based Random Forest Regression and Classification for Predictive Analytics

Predictive analysis is the procedure of extracting important patient data information in the sizable major data set. A set of enormous and intricate data is known as big-data. Predictive investigation is completed in many applications like healthcare, business, weather forecasting and thus forth, prediction investigation is performed. With big-data creation, the forecast is an important mission for discovering enhanced patient data advice from disorder data set. Several research works are done on the operation of predictive analytics. A present fuzzy principle summarization technique was designed by Mahmud et al. [1] to categorize health data. This prediction of a fuzzy version was completed centred on k fold cross validation of these data samples. It neglected to boost the forecast accuracy with big data visualization and analytics frame, and true prediction stayed a hard matter. As a way to boost the forecast accuracy with minimal time sophistication, an EGSDC-RFRC technique has been introduced.

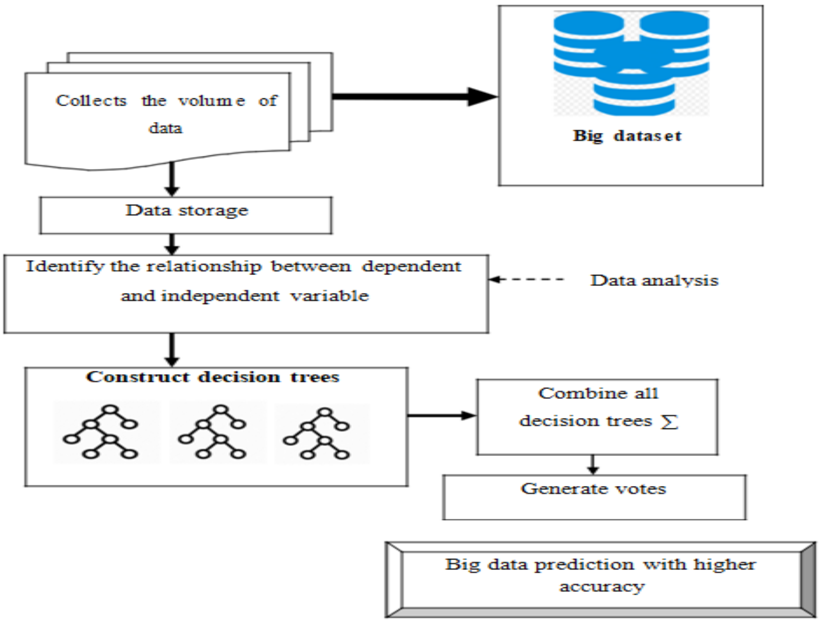
Collects the volume of data
Big dataset
Data storage
Identify the relationship between dependent and independent variable
Data analysis
Construct decision trees
Combine all decision trees ∑
Generate votes
Big data prediction with higher accuracy

Figure 1 Proposed EGSDC-RFRC Modal

Significant data set is thought of as input for patient data forecast. A massive level of data is accumulated from the data collection contemplated. Then, accumulated data was stored using Gramian symmetric matrix. The matrix assembled was at the shape of column and rows and then stored data in columns and rows matrix helped blowing risks in predictive investigation. Afterward, arbitrary decision woods ensemble learning has been introduced to its investigation of their data that was stored. Next, data investigation was completed to recognize the connection between dependent and confidential information. Bivariate Correlation was quantified to gauge data relationship. After pinpointing data, decision trees were assembled with three distinct nodes, namely, origin node, branch node and a leaf node. The tree has been assembled with the aid of training data. Decision trees were united and also voting plot has been implemented for accurate forecast benefits. The vast majority vote provided classification outcome and subsequently your mistake was calculated for true predictive investigation. Infection prediction has been attained by utilizing unstructured and ordered data. However, it neglects to reach efficient forecast with low complexity. Hence, that the BC-RFRC technique originated for improving the forecast accuracy with minimal space and time complexities.

Enhanced Gramian symmetric data collection and storage

The EGSDC-RFRC technique has already introduced an Improved Gramian Symmetric matrix for collecting big data from the data set. It's used for its storage of some massive level of accumulated data in matrix form (i.e., columns and rows). Data collection can be a highly effective way of gather extra private data information in the large data set. It's completed to predict infection probabilities. Collecting data helps analysis of procedure for extracting information and boosts the decision-making. With big-data analytics, storage space complexity is lessened in substantial data predictive investigation. [9] For disorder forecast. But data collection wasn't performed for achieving high forecast accuracy. Thus, enhanced big data collection was provided with Gramian Symmetric matrix in EGSDC-RFRC technique. Enable the range of data accumulated from the huge data set.

$$\mathbf{D1}, \boldsymbol{D_2}, \boldsymbol{D_3} \cdots\cdots \boldsymbol{D_n} \in \boldsymbol{D^1} \qquad \mathbf{(3.1)}$$

After doing information collection, information is stored in a matrix type to get information Thus, Gramian symmetric matrix is found in EGSDC-RFRC method for storage of predetermined range of information accumulated from large data set$\mathbf{D1}, \boldsymbol{D_2}, \boldsymbol{D_3} \cdots\cdots \boldsymbol{D_n}$. Here, $\boldsymbol{D^1}$ Investigation. In big-data analytics, data storage has become the main procedure. Gathered confidential data advice. Symmetric implies that almost any component is nearly Very Comparable to every. Here the accumulated information Aren't altered and saved in

$$\mathbf{g_{ij}} = \begin{bmatrix} (\boldsymbol{D1}, \mathbf{1}) & (\boldsymbol{D1}, \mathbf{2}) \cdots & (\boldsymbol{D1}, \boldsymbol{n}) \\ \vdots & \ddots & \vdots \\ (\boldsymbol{Dn}, \mathbf{1}) & (\boldsymbol{Dn}, \mathbf{2}) \cdots & (\boldsymbol{Dn}, \boldsymbol{n}) \end{bmatrix} \qquad (3.2)$$

Equation Information is inserted in to the matrix. This Allows easy accessibility and reduces gramian matrix is assembled. By Way of Example,$\mathbf{g_{ij}}$ identifies information that's saved at the very first column and row of this matrix. All of the accumulated data are organized the full-time sophistication whilst handling a huge amount of data. New columns and rows have been created if a few extra connected to quantity of columns and rows ij. By Means of saved information $\mathbf{D1}, \boldsymbol{D_2}, \boldsymbol{D_3} \cdots\cdots \boldsymbol{D_n}$.

Algorithm 1 Algorithmic process of Symmetric Data Collection based Random Forest regression and Classification

Input: Number of data $\mathbf{D}\quad_{\mathbf{1}}, \boldsymbol{D_2}, \boldsymbol{D_3} \cdots\cdots \boldsymbol{D_n}$

Output: Improve prediction accuracy with less time

```
Step 1: Begin
Step 2: Collect $\mathbf{D_i}$ from large dataset $\mathbf{D^1}$
Step 3: Store $\mathbf{D_i}$ in gramian matrix $\mathbf{g_{ij}}$
Step 4: For each row of data $\mathbf{D_i}$
Step 5: Measure the relationship $(\boldsymbol{\rho})$ between p and q
Step 6: if $\boldsymbol{\rho}$ =+1 then
Step 7: positive correlation between p and q
Step 8: else
Step 9: negative correlation between p and q
Step 10: End if
Step 11: Construct a decision tree based on relationship measure $\boldsymbol{\rho}$
Step 12: Combine decision tree $\mathbf{h_i}$
Step 13: Apply vote's v to a decision tree $\mathbf{h_i}$
Step 14: Identify majority vote for decision tree i.e. $\boldsymbol{y} = \boldsymbol{arg}\max_{\boldsymbol{n}} \boldsymbol{v}\{\boldsymbol{h_i}\}$
Step 15: Calculate generalization error
Step 16: end for
Step 17: end
```

V. Results and discussion

The experimental test of this Gramian Symmetric data-collection established Random Forest regression and Classification procedure has been ran in healthcare software using Java Language. Substantial data was believed from big data sets, namely, cardiovascular problems data set, diabetes data set and cancer data set. Heart diseases data set includes 76 features. Due to most attributes, directed into the usage of just 14 features for experimental investigation. The condition of the individual data is called through the use of 303 instances. Next, diabetes data set using 55 features was believed and these statistics were accumulated to the construction of the individual document for statistical investigation. It'd one hundred thousand instances. The accumulated patient data contained the patient level, race, sex, age, entry type, number of laboratory test conducted, HbA1c test response, diabetic medications. Likewise, breast feeding data set was employed for collecting the data for building the patient record files. The members of features were number, cell size, and contour, class features such as a benign bacterium or malicious tract therefore forth. The amount of patient data was accumulated from the significant data set and kept in a matrix in the kind of columns and rows. The preserved patient data was examined and classified as determined and independent data. The very first row represented accumulated data concerning someone and has been believed to be the practice collection. Afterward, classification was conducted for predicting the disorder types by assessing the evaluation data. 1GB patient data along with each record of 10MB has been believed for experimental procedure.

The Evaluation of the EGSDC-RFRC technique was achieved compared with different existing methods. The current processes were fuzzy rule summarization technique signalled by Mahmud et al. convolutional neural system based multimodal disease hazard prognosis (CNN-MDRP) indicated and Future Health Condition Prediction (FHCP) algorithm. The operation metrics has been evaluated on the basis of the different number of files. It's clarified in following table and graphs.

Table 1 prediction accuracy comparison

Number of files	Prediction accuracy (%)			
	fuzzy rule summarization technique	CNN-MDRP	FHCP algorithm	EGSDC-RFRC
10	50	45	56	91
20	60	55	65	86
30	65	60	72	93
40	70	64	77	90

50	76	71	83	94
60	84	76	86	94
70	83	80	89	96
80	85	82	88	94
90	83	80	86	92
100	86	84	89	96

The experimental values of forecast accuracy are displayed in table 3.1 for both EGSDC-RFRC and present techniques. The quantities of patient data files in the selection of 10 to 100 records were believed to get its behaviour the experimental job. Dining table shows a contrast of this EGSDC-RFRC procedure working with the present fuzzy principle summarization procedure. The forecast accuracy on patient data had been improved. For that reason, EGSDC-RFRC technique gained higher forecast accuracy compared to the other nation of the art procedures.

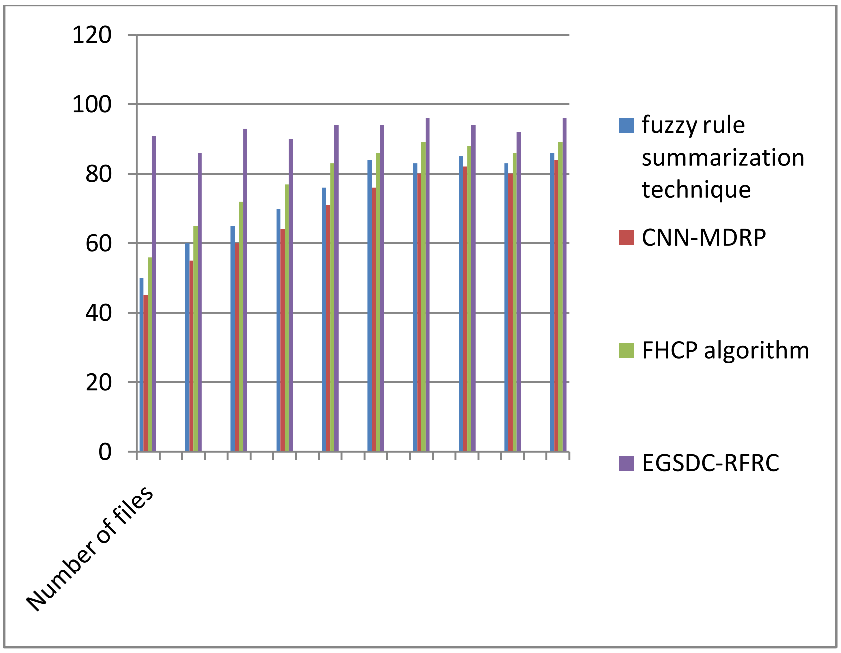

Figure 2 prediction accuracy

The experimental results reveal variations in forecast precision after a gain in the range of files. With the applying of Random woods decision tree classifier, patient data advice believed was categorized effortlessly. The classifier examined patient document between your individual data and disorder types. By utilizing Correlation, file relationship has been identified. Predicated on the significance step, your decision tree classifier was given and the disorder forecast accuracy was significantly improved. Hence, the forecast accuracy in GSDC-RFBRC technique rose by 27%, 39% and 17 percent in comparison to existing fuzzy principle summarization procedure and FHCP algorithm.

Table 2 prediction time comparison

Number of files	Prediction time (ms)			
	fuzzy rule summarization technique	CNN-MDRP	FHCP algorithm	EGSDC-RFRC
10	26	23	22	14
20	26	26	25	18
30	34	32	30	22
40	38	36	33	24
50	42	38	36	28
60	43	41	38	31
70	42	40	40	28
80	46	43	42	32
90	48	45	43	34
100	50	46	46	36

Table 2 illustrates the experimental investigation of forecast period by considering various data document. It indicates the period of time employed for prediction the disorder in the individual. Dining table shows a comparison of this consequence of forecast period for its planned and current techniques. Even the EGSDC-RFRC technique helps lower forecast time compared with all the present techniques, namely, blurred principle summarization procedure and FHCP algorithm.

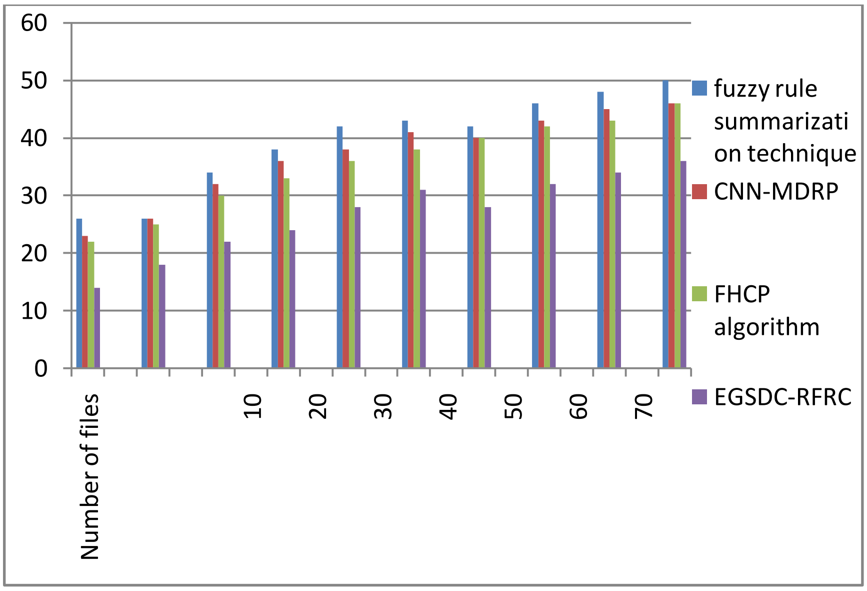

Figure 3 prediction time

Table 3 False positive rate comparison table

Number of files	False positive rate (%)			
	fuzzy rule			EGSDC-

	summarization technique	CNN-MDRP	FHCP algorithm	RFRC
10	58	50	38	10
20	42	38	30	14
30	34	30	25	13
40	30	28	26	15
50	32	29	28	18
60	34	30	27	19
70	35	29	25	17
80	34	28	27	16
90	30	26	24	18
100	32	28	26	20

Details of the experimental operation of false good speed for step by step experimentation with suggested and quitting techniques are displayed in table 3. Different quantity of files includes 10 to 100 has been known for experimental purpose and operation investigation. The suggested EGSDC-RFRC technique was compared with all existing fuzzy principle summarization procedure and FHCP algorithm. By the dining table value, its descriptive the false positive rate utilizing GSDC-RFBRC technique is paid off when comparing to other procedures.

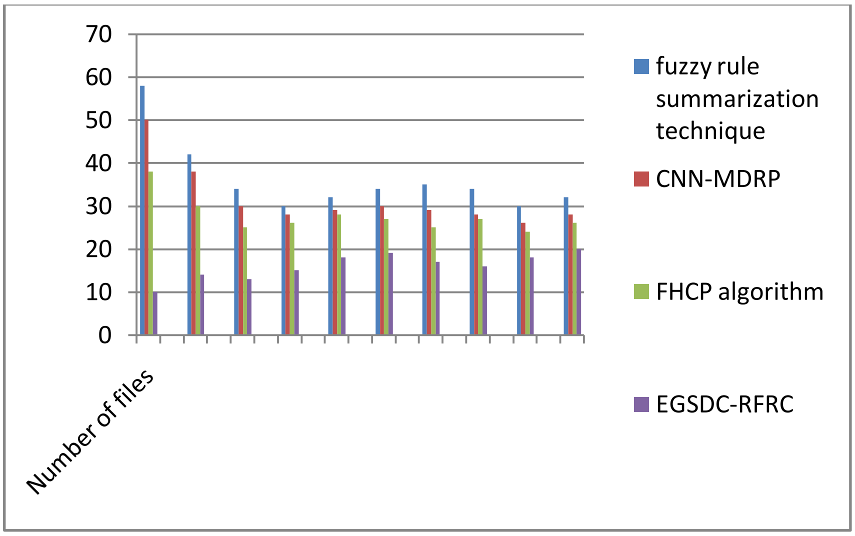

Figure 4 false positive rates

The dimension of false positive speed is displayed at Figure 4 in regards to various patient data file. The study shows the contrast of this consequence of this suggested EGSDC-RFRC procedure with the present fuzzy principle summarization procedure, FHCP algorithm. As shown in figure, the forecast of erroneous document is varied in a variety of means. But relatively, the EGSDC-RFRC technique led with minimum false positive rate when increasing the files. Afterward, the Correlation has been quantified to maintain identification of the association between independent and dependent statistics. At length, a decision tree was assembled and voting strategy has been implemented to acquire accurate disorder forecast. Along with this, outfit classifier ascertained the mistake for decreasing the false positive rate. Thus, the discovery of false positive speed has been shrunk by 51%, 45% and 39 percent in contrast to existing fuzzy principle summarization procedure and FHCP algorithm.

CONCLUSION

A reliable technique identified as EGSDC-RFRC procedure has been suggested for big health data predictive analytics. The most important intention of data that was big forecast was completed by using four distinct procedures, namely, data collection, saving data regression and data classification that was big. In the beginning, data has been accumulated from the significant data set. Then, the arbitrary woods regression and classification methods were employed for prediction of future impacts in line with the dating step. The connection was quantified using significance. The decision tree was useful for its classification of this data centred on the significance benefits. The classification provided true prognosis benefits. Experimental estimation of EGSDC-RFRC health applications was finished together with three distinct data sets, namely, heart diseases data-set, diabetes data set breast and breast cancer data collection. They plainly demonstrated the EGSDCRFRC technique improving forecast precision and reducing the prediction period, false positive rate in addition to the distance sophistication.

REFERENCES

1. Anish Jindal et.al, “Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing”, pp, Issue 99, pp: 1-1.
2. D. Pellow, D. Filippova, C. Kingsford, Improving Bloom filter performance on sequence data using k-mer Bloom filters. J. Comput. Biol. 24(6), 547–557 (2016).
3. P. Jiang, Y. Ji, X. Wang, J. Zhu, Y. Cheng, Design of a multiple Bloom filter for distributed navigation routing. IEEE Transactions on Systems Man & Cybernetics 44(2), 254–260 (2017).
4. Hong-Mei Chen et.al, 2016, “Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach”, 234 – 248.
5. J. Qian, Q. Zhu, H. Chen, Integer-granularity locality-sensitive bloom filter. IEEE Trans. Comput. 20(11), 2125–2128 (2016).
6. AbdulsalamYassine, “Mining Human Activity Patterns from Smart Home Big Data for Health Care Applications”, pp: 13131 – 13141.
7. Henry H. Chang et.al, 2009, “An Ecosystem Approach for Healthcare Services Cloud”, pp: 608 – 612.
8. Han Hu et.al 2014, “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial”, pp: 652-682.
9. J. Grimson et.al 2001, “Sharing health-care records over the Internet”, pp: 49 – 58.
10. R. Vanathi, “A Robust Architectural Framework for Big Data Stream Computing in Personal Healthcare Real Time Analytics”, pp: 97 – 104.
11.

R. Venkateswara Reddy received his B.Tech (CSE) from AITS college which is affiliated to JNTUH in 2007, M.Tech (CSE) from Hindustan University in 2011, and pursuing his Ph.D. from JJTU, Rajasthan from 2017. Presently, he is working as Assistant Professor in CMR College of Engineering &Technology,Medchal, Telangana, Hyderabad. His research interestsinclude cloud computing, data mining, and big data.

Dr. D. Murali is presently working as Professor in CMR College of Engineering &Technology,Medchal, Telangana, Hyderabad. Ph.D. in CSE from JNTU, Hyderabad in the year 2016.His areas of interest are formal language andautomata theory, digital logic design, C programming and datastructures, operating system, software engineering, compiler designing, data mining, and data warehousing.