Machine Learning Based Clinical Diagnosis of Liver Patients with Instance Replacement

J. V. D. Prasad*, A. Raghuvira Pratap and Babu Sallagundla

Department of Computer Science and Engineering, V.R. Siddhartha Engineering College, Andhra Pradesh, Vijayawada, India
E-mail: prasadjasti2018@gmail.com; raghuvirapratap@gmail.com; babunaidu.504@gmail.com
**Corresponding Author*

Received 24 June 2021; Accepted 04 August 2021;
Publication 28 October 2021

Abstract

With the rapid increase in number of clinical data and hence the prediction and analysing data becomes very difficult. With the help of various machine learning models, it becomes easy to work on these huge data. A machine learning model faces lots of challenges; one among the challenge is feature selection. In this research work, we propose a novel feature selection method based on statistical procedures to increase the performance of the machine learning model. Furthermore, we have tested the feature selection algorithm in liver disease classification dataset and the results obtained shows the efficiency of the proposed method.

Keywords: Feature selection, instance replacement, clustering.

1 Introduction

Machine learning has been a recent trend to discover knowledge in medical data that can help the clinical persons such as Doctors, Researchers, and so on. Over the years, clinical based machine learning research works has proved to uncover the hidden relationships and patterns in the medical data. However the machine learning models needs many optimizations to enhance

Journal of Mobile Multimedia, Vol. 18_2, 293–306.
doi: 10.13052/jmm1550-4646.1827

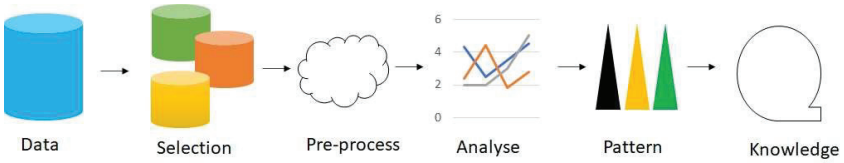

Figure 1 The five stages of KDD.

the accurate identification in the medical data. The knowledge discovery has five stages known as data selection, pre-processing, data modification, extraction of relationship and measuring evaluation. The various stages of KDD is shown in Figure 1.

Liver is the largest organ inside a human body. The main purpose of liver is to help in digestion and remove harmful substances from the body. There are three reasons for a liver disease, they are virus attacks, heavy drugs and alcohol usage and cancers. Detecting the liver disease can be done using medical imaging methods such as Sonography, CT scans [15]. These methods can cause harmful side effects. Hence much research works are introduced which are based on machine learning approach [24]. This paper focuses on proposing a new feature selection method to detect the liver disease with less cost and high performance.

Feature selection is done to reduce the dimensions of input data and to increase the speed and performance of the machine learning models [11]. Feature selection can be done in three ways namely filter [20]; wrapper [7]; and embedded [26]. Filter based method select the features based on their statistical values such as correlation. A filter-based feature selection is generally model independent, that means no matter which model is used for classification, the feature list recommendation is going to be same [8]. Filter based methods are normally used to remove the features which have less interaction with the target variable. Wrapper based feature selections are model dependent and hence the recommended feature list fully dependent on the target model [22]. If the model changes, then new set of features are recommended [6, 9]. Generally, the wrapper-based methods increase the computation time as it needs to check all the feature combinations before it recommends the best combinations [13]. Finally, the embedded methods combine both the filter based and the embedded based. Embedded methods take the advantages of other two methods [14, 16].

In this research work, we have used embedded based feature selection to find out the best possible features for each machine learning model and finally embed the results to increase the performance [1, 3].

The rest of the paper is as follows. Section 2 briefs about few existing works and Section 3 talks about the problem definition. Section 4 explains about the working of the instance replacement and the feature selection method. Section 5 compares the feature selection method with six machine learning models and finally section 6 presents the conclusion and future work.

2 Related Works

Feature selection is one of the wide techniques used in the field of machine learning for the purpose of classification [2] uses a method known as mutual information. Equation (1) shows the formula for calculating the mutual information. The h(x) represents the entropy of an input variable x and the mutual information also uses the conditional probability between x and y [31]. There are lots of errors in this method that is why many research works ignore this method. The MI works in greedy manner ignoring the overall optimal results [32].

$$MI(x,y) = H(x) - h(x|y) \tag{1}$$

One of the popular and most used method as feature selection in machine learning domain is correlation [19, 23]. Correlation gives most promising results many times. Correlation can reduce the high dimensions. There is much research [4, 12, 18] which uses correlation to identify and remove the irrelevant features from the input domain. Before the training process, the correlation can be applied to retrieve the most predominant features [33].

A hybrid feature selection method which mixes both filters based and wrapper-based approaches to obtain a good genetic FS (Feature Selection) was proposed by [10]. They have done the process of feature selection in two ways; the first way is the outer optimization where the features are selected based on the global searching methodology. After the first stage is done, an optimal sub set of features is obtained [21]. This is a wrapper based one where the authors have used mutual information. Finally, the next level of feature selection is done using filter method. The search is done locally to reduce the redundancy and error [34, 35].

Irrelevant and redundant features are removed to avoid negative impact on classification by [17]. The authors have used hybrid feature selection model based on principal component analysis and information gain. The overall training time and the number of dimensions is reduced by applying the hybrid feature selection method.

Table 1 Comparison of few related work

Reference	Area Focused	Comments
[27]	Feature Normalization	Feature normalization of non-structure features needs to be focused to improve the accuracy of the proposed classification model.
[28]	Speed of classification	The features are classified by using incremental algorithm where a proper decision should be made to select the optimal feature.
[29]	Automatic Classification	By using the semantic informations, the authors have proposed a automatic classification model. If a feature has multiple dimensions, then it is very difficult to identify the correct class.
[30]	Term Weighting	The features in the instance are read and their weights are computed based on the frequency of occurrence. Proper steps should be made on pre-processing stage to reduce the errors.
[31]	Feature Selection	The important features in the instance are identified and given more weights. This will improve the classification accuracy.

A tree-based feature selection was proposed by [25]. In their research, a tree is constructed to measure the importance of an attribute. The features are removed from the tree based on their importance. The removal process is done recursively until only the most important features are left.

3 Problem Definition

Let the available data instances D be $\{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots (X_n, Y_n)\}$, where $X_1, X_2, X_3, \ldots, X_n$ is a n-dimensional vector which represents each feature. The goal is to find M where $M \in n$ and $|M| <= |n|$.

The selected M features are highly correlated with the target class Y. The machine learning models are then trained using a new set of features $\{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots (X_n, Y_n)\}$, where $X_1, X_2, X_3, \ldots, X_n$ is a m dimensional vector which represents the selected features.

4 Clustering Cum Embedded Feature Selection Algorithm

The first task is to select five features based on embedded feature selection method for each machine learning model. The five features represent the top

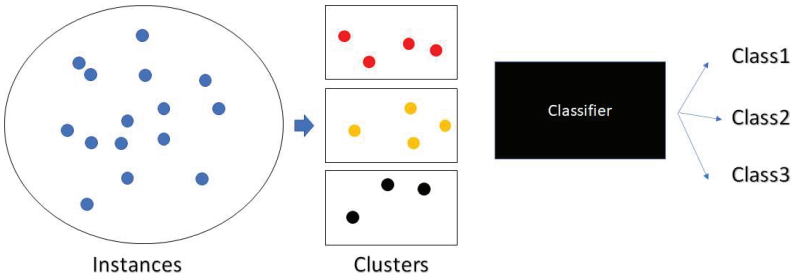

Figure 2 Architecture of the classification model.

important feature list for each machine learning model. The recommended feature list varies between each feature.

Many research works focus on weighting schemes where each feature is assigned different weights based on importance. In this work, we focus on proposing a new weighting scheme called as instance replacement where we replace a instance with another without affecting the meaning of the instance. The number of instances is going to be same and the meaning of the instance also remains as same, the only difference is the error is reduced.

Once the feature list has been built, the next step is to merge the instances into classes. We have used the k-means clustering algorithm which is let to run on each feature independently. Each value is replaced with the class id which is determined by the k-means clustering. Equation (1) shows the cost function of the clustering algorithm. Each instance is assigned to a class which has the distance minimum. Figure 2 shows the architecture of the clustering cum classification model. In the Equation (2), i represents the number of instances and j represents the current cluster and D represents the centroid of j which is found using the Equation (3). The final value j is replaced in each instance. Equation (2) talks about the distance calculation. Each point in the data space is compared with the cluster head and the total cost is the sum of all the distances. If the sum is very low, then the error rate has also reduced. Good performance can be achieved when the error rate is very less.

$$CostFunction = \sum_{j=1}^{k} \sum_{i=j} \|x_i - D_j\|^2 \tag{2}$$

$$D_j = \frac{1}{|n|} \sum_{i \in j} x(i \tag{3}$$

Table 2 Instance replacement

Feature 1	Feature 2	Feature 1 (Replaced)	Feature 2 (Replaced)
1	100	1	1
1	102	1	1
2	159	1	2
2	170	1	2
3	99	2	1

Table 3 Dataset description

Name	Mean	Std	Unique Values
Age	44.74614	16.18983	72
Gender	–	–	2
Total Bilirubin	3.298799	6.209522	113
Direct Bilirubin	1.486106	2.808498	80
Alkaline Phosphotase	290.5763	242.938	263
Alamine Aminotransferase	80.71355	182.6204	152
Aspartate Aminotransferase	109.9108	288.9185	177
Total Protiens	6.48319	1.085451	58
Albumin	3.141852	0.795519	40
Albumin and Globulin Ratio	0.947064	0.319592	73

The next issue is how to fix the number of clusters. A research work done by [5] explains a maximization function which allows the method to fix the number of classes. The number of classes varies for each feature. The algorithm uses the Equation (4) iteratively starting from 2 to 25. The optimal number of classes is fixed based on the maximum value of the equation.

An example of the instance replacement is shown at Table 1. There are two features considered where the optimal values for both the features is 2. Each instance is mapped with the corresponding class value as per minimum distance defined by Equation (4).

$$f(x) = \sum_{i=1}^{k} \pi_i N(x|\mu_i, \sum i \tag{4}$$

After the instance replacement is done, there may be lots of duplication. These duplications can be removed from the database to increase the speed of the classification process. One more important advantage of this instance

replacement is it reduces the range of values in each feature domain; hence it becomes easy for the classification model to perform the classification.

5 Experiment and Result Analysis

We have used six classifiers (SVM, kNN, NB, RF, LR and DT) to test the efficiency of the clustering cum feature selection algorithm. The dataset used for the experiment is Indian Liver Disease dataset and the description of the dataset is shown at Table 3. Linear SVM is used for the experiment. The number of trees in random forest is 7. All the nodes are split using entropy function in the decision trees. We have used Python programming language with i7 Processor, 8GB RAM.

Instance replacement is the first thing performed after normalization and other pre-processing methods. Figure 3 shows the distribution of instances after replacement. There are totally 10 input features, and the optimal k value are 11, 2, 8, 8, 5, 7, 8, 10, 6 and 5.

To test the efficiency of the proposed method, we have used four metrics Accuracy, Precision, Recall and F1 score. K-Fold validation is used (k = 10) and 70%–30% train-test split for validation purpose. Figure 4 shows the performance result for all the machine learning models.

We have found that the performance of random forest and decision tree has outperformed another machine learning model. This is because of the

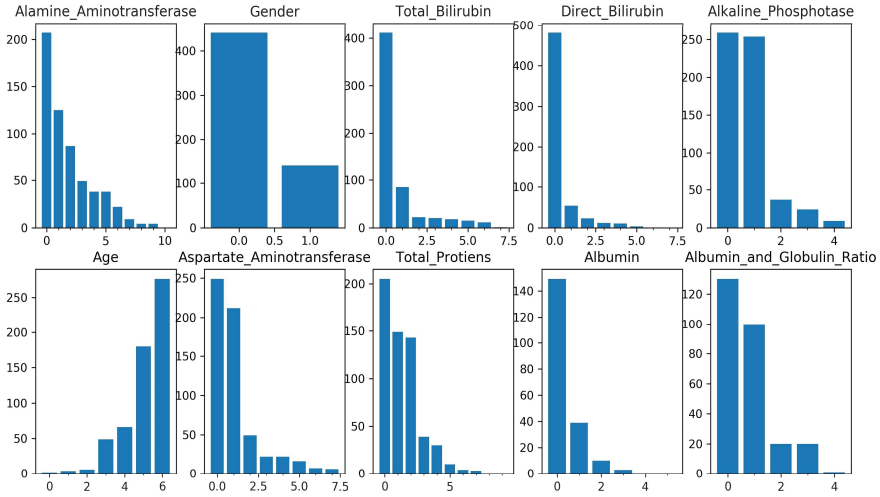

Figure 3 Distribution of instance after performing instance replacement.

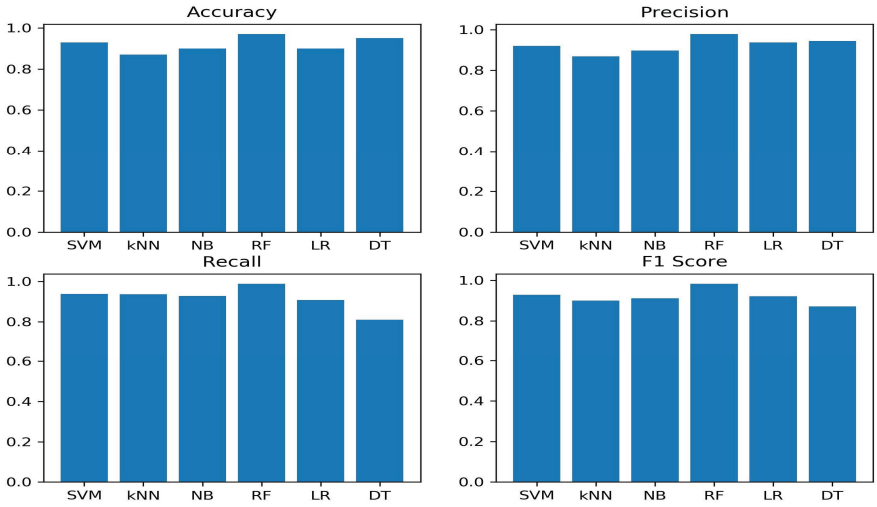

Figure 4 Performance evaluation of clustering cum FS method.

Table 4 Performance of all classifiers in 10-fold and split

	Models	Accuracy	Precision	Recall	F1-Score
k-Fold	SVM	0.93	0.92	0.93	0.92
	kNN	0.87	0.86	0.93	0.9
	NB	0.9	0.89	0.92	0.91
	RF	0.97	0.97	0.98	0.98
	LR	0.9	0.93	0.9	0.92
	DT	0.95	0.94	0.8	0.87
Spliting	SVM	0.87	0.95	0.62	0.75
	kNN	0.81	0.69	0.62	0.65
	NB	0.82	0.77	0.94	0.85
	RF	0.93	0.84	0.8	0.82
	LR	0.83	0.8	0.93	0.86
	DT	0.9	0.88	0.94	0.91

smaller number of values in each instance. This makes the branches to be easily created and the division between the instance is made very simple. The SVM classifier is also performing better because of the ability of it to work in hyperplane. Table 4 displays the performance of all classifiers.

6 Conclusion and Future Work

Considering the problem of feature selection in machine learning based classification, this paper proposes a new instance replacement strategy which reduces the redundancy and noises in the dataset. Six machine learning models were used to test the performance of the instance replacement based feature selection. Random Forest model seems to outperform all other machine learning models because of the transformed data which reduces the complexity of constructing the graph. The results fully verifies the proposed feature selection has good performance.

In the future work, we aim to analyse statistical relationship and weighting scheme to enhance the performance.

References

[1] Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M. Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, and U. Rajendra Acharya. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, 111:103346, 2019.

[2] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

[3] Zhi Peng Chang, Yan Wen Li, and Nazish Fatima. A theoretical survey on mahalanobistaguchi system. *Measurement*, 136:501–510, 2019.

[4] Jianhua Dai, Jiaolong Chen, Ye Liu, and Hu Hu. Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation. *Knowledge-Based Systems*, 207:106342, 2020.

[5] Yinlin Fu, Xiaonan Liu, Suryadipto Sarkar, and Teresa Wu. Gaussian mixture model with feature selection: An embedded approach. *Computers & Industrial Engineering*, 152:107000, 2021.

[6] Mohammad Goodarzi, Yvan Vander Heyden, and Simona Funar-Timofei. Towards better understanding of feature-selection or reduction techniques for quantitative structureactivity relationship models. *TrAC Trends in Analytical Chemistry*, 42:49–63, 2013.

[7] Adel Got, Abdelouahab Moussaoui, and Djaafar Zouache. Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach. *Expert Systems with Applications*, page 115312, 2021.

[8] Moshood A. Hambali, Tinuke O. Oladele, and Kayode S. Adewole. Microarray cancer feature selection: Review, challenges and research directions. *International Journal of Cognitive Computing in Engineering*, 1:78–97, 2020.

[9] Haouassi Hichem, Merah Elkamel, Mehdaoui Rafik, Maarouk Toufik Mesaaoud, and Chouhal Ouahiba. A new binary grasshopper optimization algorithm for feature selection problem. *Journal of King Saud University – Computer and Information Sciences*, 2019.

[10] Jinjie Huang, Yunze Cai, and Xiaoming Xu. A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters*, 28(13):1825–1844, 2007.

[11] Thirumoorthy K and Muneeswaran K. Feature selection using hybrid poor and rich optimization algorithm for text classification. *Pattern Recognition Letters*, 147:63–70, 2021.

[12] K.K. Kavitha and A. Kangaiammal. Correlation-based high distinction feature selection in digital mammogram. *Materials Today: Proceedings*, 2020.

[13] Utkarsh Mahadeo Khaire and R. Dhanalakshmi. Stability of feature selection algorithm: A review. *Journal of King Saud University – Computer and Information Sciences*, 2019.

[14] Sen Liang, Anjun Ma, Sen Yang, Yan Wang, and Qin Ma. A review of matched-pairs feature selection methods for gene expression data analysis. *Computational and Structural Biotechnology Journal*, 16:88–97, 2018.

[15] Hui Liu and Chao Chen. Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy*, 249:392–408, 2019.

[16] Bahareh Nakisa, Mohammad Naim Rastgoo, Dian Tjondronegoro, and Vinod Chandran. Evolutionary computation algorithms for feature selection of eeg-based emotion recognition using mobile sensors. *Expert Systems with Applications*, 93:143–155, 2018.

[17] Erick Odhiambo Omuya, George Onyango Okeyo, and Michael Waema Kimwele. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, 174:114765, 2021.

[18] Ashish Ranjan, Vibhav Prakash Singh, Ravi Bhusan Mishra, Anil Kumar Thakur, and Anil Kumar Singh. Sentence polarity detection using stepwise greedy correlation based feature selection and random forests: An fmri study. *Journal of Neurolinguistics*, 59:100985, 2021.

[19] Beatriz Remeseiro and Veronica Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112:103375, 2019.

[20] Sal Solorio-Fernndez, Jos Fco. Martnez-Trinidad, and J. Ariel Carrasco-Ochoa. A supervised filter feature selection method for mixed data based on spectral feature selection and information-theory redundancy analysis. *Pattern Recognition Letters*, 138:321–328, 2020.

[21] Chih-Fong Tsai, Kuen-Liang Sue, Ya-Han Hu, and Andy Chiu. Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *Journal of Business Research*, 130:200–209, 2021.

[22] Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018.

[23] Luxmi Verma, S. Srivastava, and P. Negi. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of Medical Systems*, 40:1–7, 2016.

[24] Wafaa Wardah, M.G.M. Khan, Alok Sharma, and Mahmood A. Rashid. Protein secondary structure prediction using neural networks and deep learning: A review. *Computational Biology and Chemistry*, 81:1–8, 2019.

[25] Paja Wiesaw. Tree-based generational feature selection in medical applications. *Procedia Computer Science*, 159:2172–2178, 2019. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.

[26] Yuanpeng Zhang, Shuihua Wang, Kaijian Xia, Yizhang Jiang, and Pengjiang Qian. Alzheimers disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Information Fusion*, 66:170–183, 2021.

[27] M. H. Moattar, M. M. Homayounpour and D. Zabihzadeh, "Persian Text Normalization using Classification Tree and Support Vector Machine," 2006 2nd International Conference on Information & Communication Technologies, 2006, pp. 1308–1311, doi: 10.1109/ICTTA.2006.1684569.

[28] H. Ma, X. Fan and J. Chen, "An Incremental Chinese Text Classification Algorithm Based on Quick Clustering," 2008 International Symposiums on Information Processing, 2008, pp. 308–312, doi: 10.1109/ISIP.2008.126.

[29] Lu Peng, Yibo Gao and Yiping Yang, "Automatic text classification based on knowledge tree," 2008 IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 681–684, doi: 10.1109/ICCIS.2008.4670777.
[30] M. R. Islam and M. R. Islam, "An effective term weighting method using random walk model for text classification," 2008 11th International Conference on Computer and Information Technology, 2008, pp. 411–414, doi: 10.1109/ICCITECHN.2008.4803000.
[31] Lin Lv and Yu-Shu Liu, "Research and realization of naive Bayes English text classification method based on base noun phrase identification," 2005 International Conference on Information and Communication Technology, 2005, pp. 805–812, doi: 10.1109/ITICT.2005.1609667.
[32] Ashokkumar P., Arunkumar N., Don S., Intelligent optimal route recommendation among heterogeneous objects with keywords, Computers & Electrical Engineering, Volume 68, 2018, Pages 526–535, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2018.05.004.
[33] P, Ashok K., Shiva S. G, Praveen K.R. Maddikunta, Thippa R. Gadekallu, Abdulrahman Al-Ahmari, and Mustufa H. Abidi 2020. "Location Based Business Recommendation Using Spatial Demand," *Sustainability*, 12, no. 10: 4124. https://doi.org/10.3390/su12104124
[34] Palanivinayagam, A., Nagarajan, S. An optimized iterative clustering framework for recognizing speech. Int J Speech Technol 23, 767–777 (2020). https://doi.org/10.1007/s10772-020-09728-5
[35] Palanivinayagam, A., Sasikumar, D. Drug recommendation with minimal side effects based on direct and temporal symptoms. Neural Comput & Applic 32, 10971–10978 (2020). https://doi.org/10.1007/s00521-018-3794-5.

Biographies

J. V. D. Prasad has received the M. Tech degree in Computer Science and Engineering from V.R. Siddhartha Engineering College, Vijayawada, India.

Currently pursuing Ph.D. from Department of Computer Science and Engineering from Acharya Nagarjuna University, Andhra Pradesh. His research interests include Data Mining and Parallel Computing. He has over more than 14 years of teaching experience. Currently he is working as Assistant Professor in Computer Science and Engineering at V.R. Siddhartha Engineering College, Vijayawada, India.

A. Raghuvira Pratap has received the B.Tech degree in Computer Science and Engineering from V.R. Siddhartha Engineering College, Vijayawada, India. He has received the M.Tech degree in Computer Science and Engineering from P V P Siddhartha Institute of Technology, Vijayawada, India and Currently pursuing Ph.D. from Department of Computer Science and Engineering from SRM Institute of Science and Technology, Tamil Nadu. His research interests include Machine Learning and Data analytics. He has over more than 12 years of teaching experience. Currently he is working as Assistant Professor in Computer Science and Engineering at V.R. Siddhartha Engineering College, Vijayawada, India.

Babu Sallagundla has received the B.Tech degree in Computer Science and Engineering from Priyadarsini College of Engineering, Sulurupet, India. He has received the M. Tech degree in Computer Science and Engineering from V.R. Siddhartha Engineering College, Vijayawada, India. Currently Ph.D. from Department of Computer Science and Engineering from Sarvepalli Radhakrishnan University, Bhopal, Madhya Pradesh, India. His research interests

include Machine Learning and Data Analytics. He has over more than 11 years of teaching experience. Currently he is working as Assistant Professor in Computer Science and Engineering at V.R. Siddhartha Engineering College, Vijayawada, India.