International Journal of Emerging Technologies in Engineering Research (IJETER)
Volume 9, Issue 12, December (2021)

Optimal Feature Selection and Classification Algorithms for Intrusion Detection System

Senthilnayaki B
Department of Information Science and Technology, Anna University, Chennai, Tamil Nadu, India
nayakiphd@gmail.com

Nivetha G
Department of Electronic and Communication Engineering, UCEP, Panruti, Tamil Nadu, India
gtv.evin@gmail.com

Venkatalakshmi K
Department of Electronic and Communication Engineering, UCET, Tindivanam, Tamil Nadu, India
venkata_krish@yahoo.co.in

ABSTRACT

The frequency of cyber-attacks has risen drastically as the Internet has grown in popularity, and intrusion detection systems (IDS) have become a critical component of information security. An intrusion detection system (IDS) is a piece of software that helps computer systems identify and respond to intrusions. This anomaly detection system builds a database of normal behaviour and deviations from it, which it uses to detect intrusions when they occur. Individual packets travelling over the network are monitored in network-based IDS, whereas host-based IDS investigates behaviour on a single computer or host. The feature selection contributes to the categorization time reduction. To successfully identify attacks, it was recommended and applied in this work. For this objective, optimal feature selection algorithms such as information gain ratio and genetic algorithm feature selection are applied in this proposed work. Using all these feature selection approaches, the KDD Cup dataset is utilised to find the best number of features. In addition, comparative analyses were done using different classification algorithms that were used to properly classify the data set. However, two classification algorithms are Support Vector Machine and Rule-Based Classification. This strategy is extremely good at detecting DoS attacks and reducing false alarms. The IDS can better identify assaults using the provided feature selection and classification approaches.

Index Terms – Optimal Feature Selection Techniques, data mining, classification algorithms, Intrusion detection and network security.

1. INTRODUCTION

Since computers have been networked together with a large user base, security has become a serious issue in many industries. Because of the rapid spread of internet communication and the availability of tools to enter networks, network security has become critical. Current security techniques do not effectively safeguard data stored in databases. Many other technologies, including firewalls, encryption, and authorisation systems, can provide security, but they are still subject to hacker attacks that take advantage of system flaws. To protect these systems from hackers, this project has created a unique intrusion detection system that leverages the KDD Cup data set to identify attacks using a simple feature selection technique and SVM methodology. The practise of collecting hidden predictive information from massive databases is known as data mining. It's a promising new technology that allows businesses to concentrate their data warehouses on the most critical data. Data mining may help any sort of information store. Procedures and methodologies may change when applied to different sorts of data. The internet has only lately become a part of daily life for many people. Current internet-based data processing systems are vulnerable to a variety of attacks, resulting in a variety of losses and damages. As a result, data security is becoming increasingly important. Network security's most fundamental purpose is to protect defensive networking systems from unauthorised access, use, disclosure, interruption, alteration, or destruction. Furthermore, network security reduces the risks associated with key security objectives such as confidentiality, integrity, and availability.

2. RELATED WORK

Network security has become a major concern in recent years since the emergence of the internet. There are several papers that discuss intrusion detection systems in the literature. To detect intruder attacks, intrusion detection systems (IDSs) are deployed. Sindhu et al. [1] suggested a genetic-based feature selection technique to lower the classifier's compute cost. Lee et al. [2] developed

an adaptive data mining system for intrusion detection based on association criteria and frequent occurrences acquired from audit data. Using a multiple-level hybrid classifier, Xiang and Lim [3] presented a misused IDS. Senthilnayaki et al. [4] developed a feature selection and categorization-based intrusion detection system (IDS) based on information collection. Sarasamma et al. [5] introduced a unique multilevel hierarchical Kohonen network for detecting network intrusions. To train and test the classifier, they randomly selected data points from KDD Cup 99. Jianping Li et al. [6] introduced a new approach for choosing relevant feature sets for network intrusion detection based on the Continuous Random Function. In the IDS literature, there are several classification methods based on SVM. For effectively categorising data, Snehal A. Mulay et al. [7] suggested a Tree Structured Multiclass SVM method. Reprocessing is discussed in a number of papers [8] [9]. Instead of computing them accurately at the cost of reduced performance, time, and space complexity, most real-world situations necessitate an optimum and acceptable solution. The feature selection procedure began with either a null set of features that were added one at a time or a complete set of features that were removed one at a time. For the creation of an IDS, Li et al. [11] developed a wrapper-based feature selection technique. Senthilnayaki et al. [10] provide a feature selection approach to handle the statistical strategy for analysing the large KDD Cup dataset.

Many books and articles on categorization techniques and tools may be found in the literature. [14]. Support vector machines (SVMs) are classifiers designed specifically for binary classification. For IDS, Debar et al. [12] created a Neural Network (NN) model. By incorporating membership into each data item, Du Hongle et al [13] suggested an enhanced v-FSVM. Senthilnayaki et al. [15] proposed a new network intrusion detection learning approach based on a fuzzy rough set feature selection and a modified KNN classifier algorithm that identifies effective attributes from the training dataset, calculates rules based on the best attribute values, and correctly classifies all examples in the training and testing datasets. [16] An intrusion detection system based on SVM was presented.

3. PROPOSED SYSTEM IMPLEMENTATION AND RESULTS

The data gathering agent collects information from the KDD '99 cup data collection. This data is sent to the data preparation module, which preprocesses it. The data from the KDD Cup dataset might be legitimate or malicious. Preprocessing techniques are essential for data minimization since processing large amounts of network traffic data with all the attributes necessary to detect attackers in real time and give preventative strategies is difficult. The suggested system architecture is depicted in Figure 1 and includes preprocessing, a feature selection module, and a classification module. The preprocessing module removes unwanted data, redundant records, and missing values from the collected KDD Cup dataset. Next it is given to the feature selection module, which reduces the number of important features from the total feature. Due to this, it obtains less processing time. Finally, the classification module is to classify selected records into normal and attack categories.

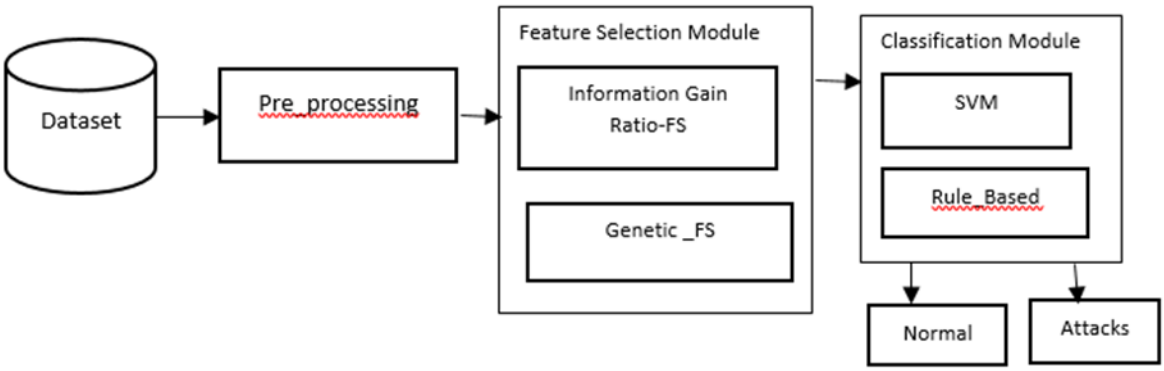

Figure 1 Intrusion Detection System Architecture

The use of rules triggered by intelligent agents and fired using the rule system enhances judgements on anomalous intrusion detection and prevention in this system. The main advantage of using rules with a knowledge base is that it makes it easier to make good intrusion judgments. The SVM is a learning machine for binary classification and regression estimation. Due to two fundamental qualities, they are becoming increasingly intriguing as a new paradigm of categorization and learning. The Information Gain Ratio was used to create this attribute selection strategy. To do this, the data set D is partitioned into n classes,

Ci. The agent chooses the Fi features with the most non-zero values, and the Information Gain Ratio (IGR) is determined using the formulae [4] below:

$$\text{Info (D)} = -\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|D|}\right] log_2\left[\frac{freq(C_j,D)}{|D|}\right] \quad (1)$$

$$\text{Info (F)} = \sum_{i=1}^{n} \left[\frac{|F_i|}{|F|}\right] * info\ (F_i) \quad (2)$$

$$\text{IGR } (A_i) = \left[\frac{Info\ (D) - Info\ (F)}{Info\ (D) + Info\ (F)}\right] * 100 \quad (3)$$

An Algorithm for Information Gain Ratio:

Input: KDD Cup Dataset with 41 features

Output: Selected Feature (10)

Step 1: Gather the non-varying values in each column.

Step 2: Determine the frequency of each column.

Step 3: Using equation 1, calculate information info (D).

Step 4: Using equation 2, calculate info (f).

Step 5: Determine the information gain ratio equation 3

Step 6: Using information gain values, select important features.

Table 1 shows that different types of attacks are in the KDD Cup dataset. It shows four major types, such as probing, denial of service (DoS), user to root (U2R), and remote to user (R2L). Table 2 shows all features in KDD, 99 datasets.

Table 1. Different types of attacks in KDD dataset.

Attack Classes	Attacks
Probing	ipsweep, nmap, portsweep, satan
Denial of Service (DOS)	back, land, neptune, pod, smurf, teardrop
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Remote to User (R2L)	ftp_write,guess_passwd,imap,multihop,phf,spy, warezclient, warezmaster

Table 2 All Features in KDD'99 dataset

S. No	Feature Name	S. No	Feature Name
1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

A rule-based classification algorithm is used to categorise various types of attacks. Figure 2 displays the results of the performance analysis in terms of accuracy, using the Rule-Based Classifier to categorise the attacks. The detection accuracy is obtained for total features, selected features using information gain ratio and selected features using a genetic algorithm, as shown in figure 2. Table 3 shows the list of selected features using a genetic algorithm.

Table 3 List of Selected Features using GA

S.No	S.No from KDD Dataset	Selected Feature
1	2	Protocol type
2	3	Service
3	4	Src_byte
4	5	Dst_byte
5	6	Flag
6	27	Diff_srv_rate
7	33	Dst_host_srv_count
8	40	Dst_host_rerror_rate
9	41	Dst_hostsrv_rerror_rate

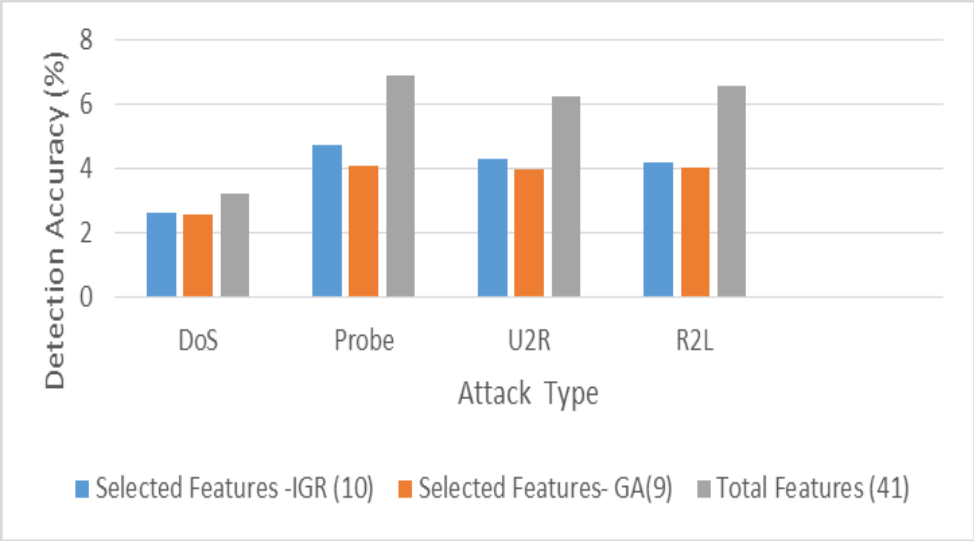

Figure 2 Performance analysis for Rule based Classification

Figure 3 shows an SVM-based classification algorithm being used to categorise various types of attacks. It displays the results of the performance analysis in terms of accuracy, using the SVM-based classifier to categorise the attacks. The detection accuracy is obtained for total features, selected features using information gain ratio and selected features using a genetic algorithm, as shown in figure 3.

Table 4 illustrates the time analysis for DoS attacks for different features dataset assaults using SVM classification for the 5000 data collected. It is observed that selected features take less time for the computation process compared with all other features.

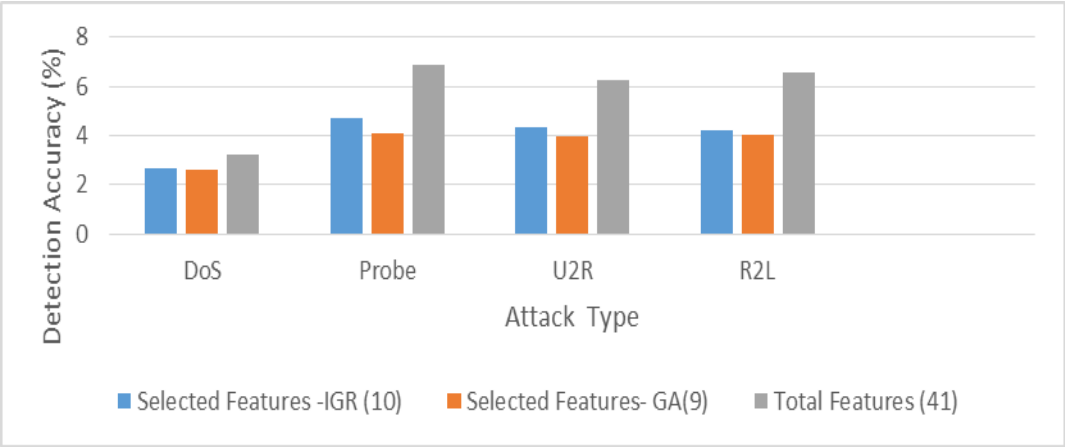

Figure 3 Performance analysis for SVM

Table 4 Time Analysis for DoS attack in SVM

Exp No.	Time in (ms)		
	Selected features using OFS (10)	Selected features using GA (9)	Total features (41)
1	2.59	2.48	3.31
2	2.75	2.72	3.26
3	2.85	2.79	3.23
4	2.65	2.64	3.18
5	2.45	2.35	3.17
Avg	2.65	2.59	3.23

Table 5 illustrates the time analysis for probe attacks for different features dataset assaults using SVM classification for the 5000 data collected. It is observed that selected features take less time for the computation process compared with all other features.

Table 5 Time Analysis for probe attack in SVM

Exp No.	Time in (ms)		
	Selected features using OFS (10)	Selected features using GA (9)	Total features (41)
1	5.37	4.35	7.25
2	5.01	4.15	6.91
3	4.99	4.01	6.87
4	4.15	4.08	6.78
5	4.17	4.03	6.75
Avg	4.73	4.12	6.91

Table 6 illustrates the time analysis for U2R attacks for different features of dataset assaults using SVM classification for the 5000 data collected. It is observed that selected features take less time for the computation process compared with all other features.

Table 6 Time Analysis for U2R attack using SVM

Exp No.	Time in (ms)		
	Selected features using OFS (10)	Selected features using GA (9)	Total features (41)
1	4.51	4.05	6.52
2	4.41	3.99	6.25
3	4.37	4.17	6.23
4	4.25	3.88	6.17
5	4.19	3.89	6.07
Avg	4.34	3.99	6.24

Table 7 illustrates the time analysis for R2L attacks for different features of dataset assaults using SVM classification for the 5000 data collected. It is observed that selected features take less time for the computation process compared with all other features.

Table 7 Time analysis for R2L attack using SVM

Exp No.	Time in (ms)		
	Selected features using OFS (10)	Selected features using GA (9)	Total features (41)
1	4.55	4.15	6.95
2	4.25	4.03	6.55
3	4.26	4.15	6.87
4	4.1	4.02	6.45
5	3.99	3.99	6.1
Avg	4.23	4.06	6.58

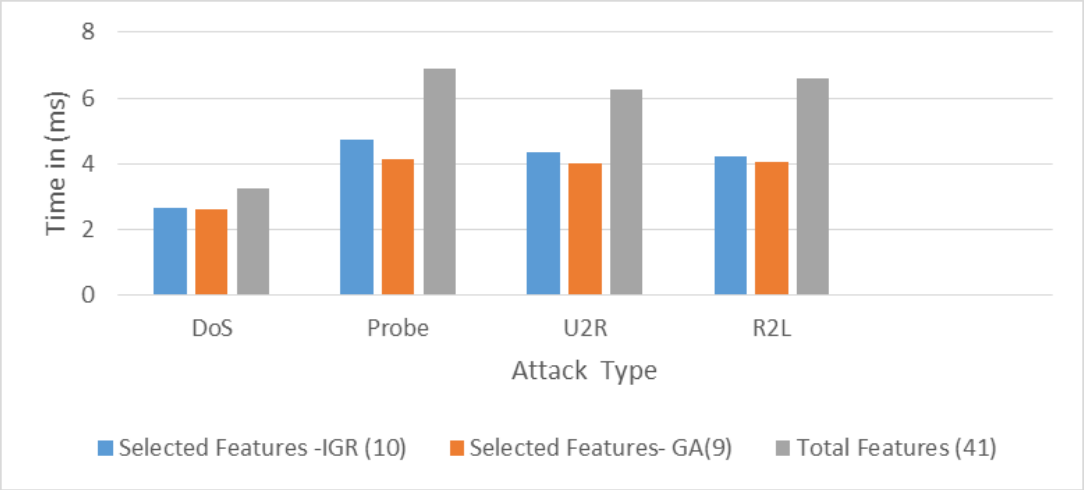

Figure 4 Average Computation time for all attacks using SVM

The performance accuracy analysis for SVM for various sizes of data is shown in Table 8. As the number of records increases, the accuracy of detecting records with specific characteristics using GA improves, eventually approaching that of recognising records with all attributes using SVM.

Table 8 Accuracy Analysis using SVM

S.No	Records in Experiment	DoS	Probe	U2R	R2L
1	5500	99.15	92.05	91.58	91.57
2	10000	99.25	94.27	93.87	93.85
3	13000	99.31	94.45	93.45	93.78
4	17000	99.41	93.99	93.87	93.14
5	25000	99.17	93.99	94.12	93.75

This proposed work looks at the installation and results of the recommended system. The outcome of this work, as well as recommendations for future efforts, will be presented in the next section.

4. CONCLUSION

By combining an Optimal Feature Selection approach with classification algorithms, a new intrusion detection model for system security is developed and implemented in this proposed work. Recognizing and categorising records using all forty-one features of the KDD '99 cup data set takes a lengthy time, it is discovered. The suggested feature selection method picks just the most significant qualities, cutting down on the time it takes to identify and categorise data. SVM and genetic-based feature selection also aid in improving accuracy. The proposed IDS's key benefit is that it minimises false positive rates while simultaneously reducing calculation time.

REFERENCES

[1] Sindhu, S, Geetha, S. and Kannan, A. “Decision Tree based Light Weight Intrusion Detection using a Wrapper Approach”, Expert Systems with Applications, Vol. 39, pp. 129–141, 2012.
[2] Farid D.M, Jerome Dormont, NouriaHarbi, Nguyen HuuHoa and Rahman, M.Z. “Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification”, International Conference on Computer Systems Engineering, Version 1, pp. 321-337, 2010.
[3] Wei Wang, Xiangliang Zhang, Sylvain Gombault and Svein J. Knapskog, “Attribute Normalization in Network Intrusion Detection”, 10th International Symposium on Pervasive Systems Algorithms and Networks, pp. 543-559, 2009.
[4] Senthilnayaki Balakrishnan, Venkatalakshmi, K and Kannan, A ‘Intrusion Detection System Using Feature Selection and Classification Technique’, International Journal of Computer Science and Application, vol.3, no.4, pp.146-151, 2014.
[5] SarasammaS., Zhu, Q. and Huff, J. “Hierarchical Kohonen Net for Anomaly Detection in Network Security”, IEEE Transactions on System, Man,Cybernetics, Part B,Cybernetics, Vol. 35, No. 2, pp. 302-312, 2005.
[6] Wang Jianping, Chen Min and Wu Xianwen, “A Novel Network Attack Audit System based on Multi-Agent Technology”, Physics Procedia, Elsevier,Vol. 25,pp. 2152 – 2157, 2012.
[7] Snehal A. Mulay, Devale, P.R. and Garje, G.V. “Intrusion Detection System using Support Vector Machine and Decision Tree”, International Journal of Computer Applications, Vol.3, pp.975-987,2010.
[8] DaramolaO.Abosede, AdetunmbiA.Olusola, AdeolaS.Oladele,. “Analysis of KDD’99 Intrusion Detection Dataset for Selection of Relevance Features”,Proceedings of the World Congress on Engineering and ComputerScience ,Vol.I , October 20-22, 2010.
[9] Senthilnayaki, B., Chandralekha, M., & Venkatalakshmi, K. (2015). Survey of data mining technique used for intrusion detection. Int. J. Technol, 7(2), 166-171.
[10] Leng J, Valli C, and Armstrong L. “A Wrapper-based Feature Selection for Analysis Large Data Set”, Proceedings of 2010 3rd International Conference on and Electrical Engineering (ICCEE), pp. 167-170, 2010.
[11] Senthilnayaki B, Venkatalakshmi K and Kannan A 2015, “Intrusion Detection Using Optimal Genetic Feature Selection and SVM based Classifier”, 3rd International Conference on Signal Processing, Communication and Networking (ICSCN). Pp 1-4, 2015.
[12] Devale.P.R ,Garje.G.V., SnehalA.Mulay, 2012. “Intrusion Detection System using Support Vector Machine and Decision Tree “, International Journal of Computer (0975 – 8887), Vol. 3, June 2010.
[13] Debar, H., Becker, M. and Siboni, D. “A Neural Network Component for an Intrusion Detection System”, IEEE Symposium on Research in Computer Security and Privacy, pp. 240-250, 1992.
[14] Du Hongle, TengShaohua and Zhu Qingfang, “Intrusion detection Based on Fuzzy support vector machines”, International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 639-642, 2009.
[15] Senthilnayaki, B., Venkatalakshmi, K., & Kannan, A. (2019). Intrusion detection system using fuzzy rough set feature selection and modified KNN classifier. Int. Arab J. Inf. Technol., 16(4), 746-753, 2019.
[16] Senthilnayaki, B., Chandralekha, M., & Venkatalakshmi, K. (2015). Survey of data mining technique used for intrusion detection. International Journal of Technology, 7(2), 166-171.

Authors

Balakrishnan Senthilnayaki has completed MTech and PhD at (CEG) Anna University, Chennai - 25. She has 10 years of teaching experience. Currently, she is working as a Teaching Fellow of the (CEG) Anna University, Chennai. She has 17 publications in journals and conference proceedings. Her areas of interest include Data Mining, Machine Learning and Soft Computing.

Krishnan Venkatalakshmi has completed ME and PhD at Thigarajar Engineering College, Madura i. She has 20 years of teaching experience. Currently, she is head and Assistant Professor in the Department of Electronics and Communication Engineering at Anna University (UCET) Tindivanam. She has more than 75 publications in reputed journals and conference proceedings. Her area of interest includes Signal Processing, VLSI, Wireless Networks, Wireless Communication and Instrumentation.

Nivetha Gopalakrishnan has completed her M.E at Periyar Maniammai College of Technology for Women,Vallam and Ph.D at Anna University, Chennai - 25. She has 15 years of teaching experience. Currently, she is working as a Assistant Professor in the Department of Electronics and Communication Engineering at Anna University (UCEP) Panruti. She has more than 15 publications in reputed journals and conference proceedings. Her area of interest includes Image Processing, Wireless Sensor Networks and Data Mining.