International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Naïve Bayes guided Binary Firefly Algorithm for Gene Selection in Cancer Classification

R.R.Rajalaxmi, E.Gothai, .R.Thamilselvan, P.Gokila Bindha, P.Natesan

Abstract— The bioinformatics research must deal with the analysis of the large volume of data. Disease classification deals with the identification of relevant genes in almost all gene expression analyses, where researchers attempt to select a minimum number of genes with exceptional performance. The gene selection process mainly selects significant genes related to the disease. This work aims to accomplish relevant genes from large volume of candidate genes that help to identify cancers. In the proposed work, Binary Firefly Algorithm (BFA) helps to identify related genes using the Naive Bayes classifier. Based on the experimental results, Naïve Bayes guided Binary Firefly Algorithm (NBBFA) provided high accuracy with fewer genes

Keywords: Microarray data; Gene Selection; Naïve Bayes classification; Binary Firefly Algorithm..

I. INTRODUCTION

Microarray technology empowers to measure expression levels of thousands of genes. The microarray data contain information about diagnostic, prognostic, and biological details. Informative genes identification in cancer classification is a new problem [1][14]. The gene selection problem is a challenging task due to a greater number of genes while comparing with the less quantity of samples, noisy and unrelated data [19] present in the microarray data. Gene selection method chooses gene subsets that improve the classification accuracy of the samples. Figure 1 depicts the diagrammatic representation of gene selection.

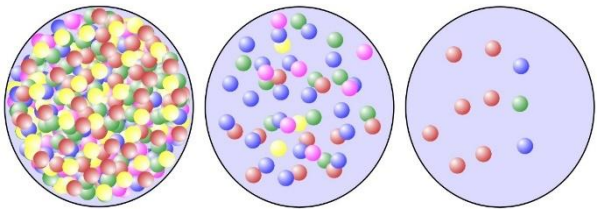

Figure 1 Gene Selection

Gene selection involves three approaches, namely the wrapper, the filter, and the embedded approaches. The filter method employs statistical procedures to find the properties of genes to identify relevant genes. These methods mainly perform the gene selection process considering as a pre-processing step without any classification algorithm. Wrappers utilize machine learning approaches to search for the best genes among all gene subsets. In the embedded method, it performs the gene selection process while constructing the model. In contrast to filter and wrapper approaches, in embedded methods, the learning step and the gene selection step are combined – i.e., the function structure used in the method plays a vital role.

Informative genes identification from thousands of genes is a difficult task. There are many methods such as filter method, wrapper method, and embedded methods to select relevant genes from microarray data. These methods never consider the correlation between the genes in the selection procedure and suffer from excessive computational complexity. To get a better efficiency Particle Swarm Optimization (PSO) and other methods are used. In this work, the Binary Firefly Algorithm (BFA) is used to improve accuracy.

II LITERATURE REVIEW

Ben-Dor et al. [2] studied three different sets of gene expression data measured across tumor(s) and standard samples: In the first dataset, it contains 2,000 genes with 62 epithelial colon samples. The second dataset consists of approximately 100,000 genes with 32 samples. The third dataset contains around 7,100 genes with 72 examples. They employed scoring methods to measure the separation of tissue type with individual gene expression levels. Followed by this, classification methods are used to measure the classification ability of whole expression profiles. They performed experiments by employing nearest neighbor classifiers, Support Vector Machine (SVM), AdaBoost, and a novel cluster-based classification method along with Leave-One-Out Cross-Validation (LOOCV) on the datasets.

Manuscript published on November 30, 2019.
* Correspondence Author
Dr.R.R.Rajalaxmi*, Department of Computer Science and Engineering, Kongu Engineering College, Perunduai, Tamil Nadu, India. (Email: rrr@kongu.ac.in)
Dr.E.Gothai, Department of Computer Science and Engineering, Kongu Engineering College, Perunduai Tamil Nadu, India. (Email: egothai@kongu.ac.in)
Dr.R.Thamilselvan, Department of Computer Science and Engineering, Kongu Engineering College, Perunduai Tamil Nadu, India. (Email: r_thamilselvan@kongu.ac.in)
Ms.P.Gokila Bindha, Department of Computer Science and Engineering, Kongu Engineering College, Perunduai Tamil Nadu, India.(Email: brindha@kongu.ac.in)
Dr.P.Natesan, Department of Computer Science and Engineering, Kongu Engineering College, Perunduai , Tamil Nadu, India. (Email: natesanp@kongu.ac.in)

Li et al. [7] suggested a multivariate method to identify a subset of relevant genes for sample classification. The algorithm used colon and leukemia data sets. Genetic Algorithm (GA) amalgamated with the k-Nearest Neighbour (kNN) method identify genes to discriminate between normal and tumor. The identified genes are further used to classify the test set samples. The proposed method is robust to deal with a sizeable noisy dataset for classification. Also, a multivariate analysis is employed to capture the correlated relationship among the data. Troyanskaya et al. [16] also addressed the problem of the identification of relevant genes from microarray data. The work systematically evaluated the performance of each method with simulated and biological data of varying noise levels. All methods identify a significant fraction of the informative genes with noise levels like that of actual data.

He et al. [5] suggested Support Vector Machines (SVMs) with a novel regularization method to detect essential genes for cancer classification. An individual nonconvex penalty is levied over the loss function in the SVM. The proposed approach removes redundant genes automatically with useful classification.

Li et al. [8] developed PSO with GA and used SVM for gene selection. In this method leukemia, colon cancer, and breast cancer datasets are used for validation. Experimental results exhibit higher classification accuracy compared to the other feature selection methods.

A combined PSO and Tabu Search (HPSOTS) approach for gene selection is developed by Shen et al. [15]. Tabu Search (TS) is used to overcome local optima issue. Three different microarray datasets are used for experimental analysis. The results of HPSOTS are compared with stepwise selection, PSO, and TS algorithm. The experimental results indicate that HPSOTS is effective in gene selection and mining high dimension data.

Chuang et al. [3] developed Improved Binary Particle Swarm Optimization (IBPSO) with the kNN method for subset evaluation in gene selection. IBPSO has been tested on the leukemia microarray data set. Experimental results reveal that IBPSO's performance was superior to the simple binary PSO and other related works. IBPSO updated particle position with a modified rule, which leads to identifying a minimal number of genes each iteration, and finally yields an optimal subset of genes with improved classification accuracy.

Mohamad et al. [9] presented an improved PSO integrated with cross-validation (LOOCV) to identify a subset of informative genes for cancer classification. As per the experimental results, IPSO yielded good accuracy with the small number of genes compared to other work.

Alba et al. [1] designed a geometric PSO algorithm to deal with high-dimensional microarray data. The proposed approach, called Speed-constrained Multi-objective PSO (SMPSO) handles the high-velocity problem in particle position update when the velocity becomes too high. The proposed method is compared concerning five multi-objective metaheuristics algorithms most famous in the area. The experiment results show that SMPSO attains outstanding results in terms of both speed and accuracy.

Novel Hybrid Framework (NHF) for gene selection is [19] developed for high dimensional microarray data. This framework combined the F-score, Information Gain (IG), GA, PSO, and SVM. The results are compared with PSO, Ant Colony Optimization (ACO), GA, and Simulated Annealing (SA) methods on five standard microarray data. Performance results exhibit that NHF identifies informative gene subsets from noisy data set along with the correlated structure of the data. Also, NHF produces better prediction accuracy with a smaller subset of features than the other methods.

A recursive PSO based wrapper method for gene selection [10] integrated with various filter-based ranking strategies exhibits considerable improvement in the classification accuracy. Experiments were performed with five freely accessible standard microarray datasets. The recursive PSO method chooses a minimum number of genes with the best accuracy in all the cases.

Spider Monkey Optimization algorithm [11] is used to identify the number of genes in cancer data. Experimental performed with various benchmark cancer datasets reveals that it outperforms other methods with the minimum amount of genes and maximum classification accuracy. The authors [17] developed Bacterial Colony Optimization for gene selection. Here, the population contains multi-dimensionality variables to select gene subsets. Also, the proposed method can select a robust genes subset to attain improved classification accuracy.

In [13], a framework (C-HMOSHSSA) for gene selection using the salp swarm algorithm (SSA) and multi-objective spotted hyena optimizer (MOSHO) was developed. The proposed algorithm hybridizes the features of both SSA and MOSHO to facilitate its exploitation and exploration capability. The experimental results illustrate that the proposed procedure significantly outperforms compared with other popular methods. Also, it significantly identified new sets of useful and biologically relevant genes.

Thomas et al. [15] developed a statistical regression modeling approach to determine genes that are differentially expressed among two predefined sample collections in DNA microarray experiments. The model is created with valid assumptions, uses appropriate statistical measures to represent heterogeneous and gnomically complex data. This methodology effectively-identified 141 differentially expressed gene subset from AML and ALL data with a 1% genomic significance level.

With the help of this method, the work identified a set of genes with their expression profiles correlated with thrombopoietin and noticed that genes whose expression associated with AML treatment outcome lie in recurrent chromosomal location compared among different sample groups within the AML samples. Many techniques are used to obtain better accuracy and the minimal number of genes. To achieve improved accuracy BFA combined with Naive Bayes classifier is proposed in this work.

III PROPOSED WORK

In this work, Naïve Bayes guided Gene selection using binary firefly algorithm is proposed.

A. *Firefly Algorithm*

The firefly algorithm (FA) is [18] created on the flashing behavior of fireflies. The firefly attracts other fireflies using the flash as a signal system. The following rules govern the interaction between fireflies:

- All the fireflies are unisex, and their attraction with other fireflies is independent of their sex.
- The brightness of a firefly is associated with its attractiveness. However, the brightness intensity of a firefly supposed by another firefly is not directly proportional to the distance between them. Thus, the brighter firefly attracts the dimmer firefly. If none of the fireflies is brighter than a given firefly, then it moves randomly.
- The brightness of a firefly corresponds to the value of an objective function to optimize.

Most of the fireflies produce quick and periodic flashes. The flashes follow a specific pattern for a precise species. A kind of flash pattern is formed to attract both the males and females to each other, as shown in Figure 1. Females of a group respond to an individual pattern of the same group of the male. The strength of light at a particular distance r from the light source follows the inverse square law. The intensity of light becomes weaker when the distance increases. So, most fireflies have visibility up to a few hundred meters at night, which enables them to communicate with each other.

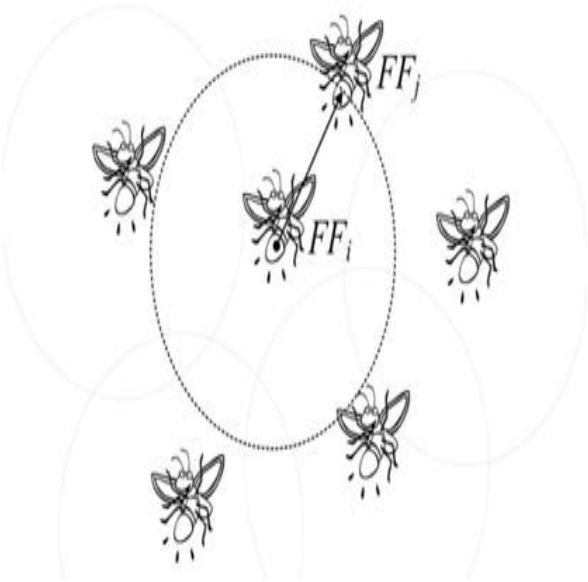

Figure 1 Behavior of fireflies

The main characteristics of the firefly algorithm are:

Brightness: This mainly depends on the objective function. For simpler optimization problems, the firefly (x) brightness is abridged to the objective function for x :

$$I(x)=f(x) \quad (1)$$

Attractiveness: It is proportionate to the brightness as supposed by the other neighbors. Generally, the attractiveness function monotonically decreases by the actual distance r. A general representation of it is:

$$\beta = \beta_0 e^{-\gamma r2} \quad (2)$$

where r denotes the distance between two fireflies, β_0 depicts the attractiveness for r = 0, and γ represents the bright absorption coefficient (a constant).

Distance: It measures the Euclidean distance among the two fireflies x_i and x_j

$$r_{ij} = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})2} \quad (3)$$

where x_{ik} stands for the kth component of the i^{th} firefly

Movement: The firefly i moves to new position as attracted by another brighter firefly j is determined by:

$$x_{i+1}=x_i+\beta_0 e^{-\gamma r_2}(x_j-x_i)2+\alpha(rand-0.5) \quad (4)$$

The first and the second terms relate to the attractiveness. The third term is the randomization.

Figure 3 gives the algorithm pseudo-code.

```
Input:
Generate initial population of fireflies n within d-dimension
i.e. x_ik , i = 1, 2, . . . , n and k = 1, 2, . . . , d
Evaluate the fitness of each firefly f (x_ik)
Output:
Obtained best solution: x,min
begin
  repeat
    for i = 1 to n
      for j = 1 to n
        if (Ij < Ii)
          Move firefly i to j using Equation (4)
        end if
        Vary attractiveness with distance r via exp[−r2]
        Evaluate new solutions and update light intensity
         using Equation (1)
      end for j
    end for i
    Rank the fireflies and choose the current best
    until (stop condition)
end
```

Figure 2 Firefly algorithm

The initial positions of the fireflies in the algorithm are unknown. So, it is randomly initialized. The algorithm tends to find the optimal position of fireflies. Then the population is initialized by producing random numbers with dimension d and number N in the search space xij. Then the initialized population is evaluated using fitness function. The movement of fireflies is influenced by the fitness values of all fireflies. The dimmer firefly i move towards, the brighter firefly j as per equation (4)

B. *Binary Firefly Algorithm*

In the firefly algorithm, brightness of fireflies represents their positions. For a binary firefly algorithm, 0s and 1s indicate the position of fireflies [12], as shown in Table 1. In gene selection problem, we consider each firefly is represented as a binary vector with d number of genes denoted as $g_i = (g_{i1}, g_{i2}, g_{i3}, \ldots. g_{id})$, $i = 1,\ldots,n$ where 'n' denotes the number of fireflies. Each element of g_i is either 0 or 1, indicating absence or presence of a gene. i.e. 1 shows a selected gene, and 0 indicates a non-selected gene.

Table 1 Binary representation of firefly

FIRE FLY	GENES						
	G 1	G 2	G 3	G 4	G 5	G_j	G_n
F_i	1	0	0	0	1	1 ……..	1

The position of firefly shifts between ‘‘1’’ and ‘‘0’’ in discrete space. To accomplish this, a transfer function maps brightness values to probability values for updating the positions, i.e., a transfer function changes a position vector’s values from 1 to 0 and vice versa. Thus, it makes the fireflies to move in a binary space. The most commonly employed transfer function is Sigmoid function as per the following equation:

$$S(x_i^k(t))=\frac{1}{1+e^{-x_i^k(t)}} \quad (5)$$

$$x_i^k(t+1)=\begin{cases}0 & If\ rand < S(x_i^k(t))\\ 1 & If\ rand > S(x_i^k(t))\end{cases}$$

where x_i^k designate the position of i[th] firefly at iteration t in k[th] dimension.

Each firefly moves find the optimal gene set from the search space for cancer identification based on the fitness evaluation done through the Naive Bayes classifier. In machine learning, naive Bayes classifier is a simple probabilistic classifier using Bayes theorem [4] based on class conditional independence among the features. It requires a few parameters and highly scalable for a learning problem. Classification accuracy is used for evaluation and calculated as follows:

Accuracy= (# correctly classified samples)/(#of samples) (6)

Input:
Generate an initial population of fireflies n with d-dimension i.e. x_{ik}, i = 1, 2, . . . , n and k = 1, 2, . . . , d
Output: Number of relevant genes and accuracy
Begin
  Initialize the population
  Evaluate the fitness value using the Equation 6
  while (max iteration is not met)
    for i = 1 to n
      for j = 1 to d
        update the positions of firefly using the Equation 4
        Change the positions using sigmoid function as per Equation 5
        if rand < sigmoid(x_i^{old})
          x_i^{new}=0
        else
          x_i^{new} =1
        end if
      end for
      Evaluate the fitness of new solutions by Equation 6
    end for
    Rank the fireflies and detect the current best solution
    end while
  Find the best solution
end

Figure 3 Pseudocode of BFANB

Finally, the firefly with the best fitness value is selected, and the corresponding gene subset is determined. Figure 3 describes the pseudocode of Gene Selection using the Binary Firefly Algorithm.

IV RESULTS AND DISCUSSION

In this section, the experimental results of the NBBFA with microarray datasets are compared with PSODT [6]. The Parameter values for constants used in BFA are:

alpha=0.2
gamma=1.0
delta=0.97
Number of iterations=100
Number of fireflies=50

Microarray datasets with details such as classes, features, and classes size used in this work are given in Table 2

Table 2 Dataset details

Dataset	Feature size	Sample size	Class size
Lung cancer	12601	203	3
Brain_Tumor1	5921	90	9
Brain_Tumor2	10368	50	11
Leukemia1	5328	72	3
Leukemia2	11226	72	3
SRBCT	2309	83	2
Prostate_Tumor	10510	102	2

To study the effectiveness of the proposed NBBFA, a comparison with other popular classification algorithms like SVM, BPNN, CART, and PSODT is performed. Table 3 depicts the comparison of the proposed method with the different four algorithms. NBBFA is superior to the others, except Lung cancer and Leukimia2, even though it identifies minimal genes.

Table 3 Classification Accuracy of the datasets

Dataset	SVM	BPNN	CART	PSODT	NBBFA
Lung cancer	97.70	39.07	73.61	100	98.55
Brain_Tumor1	49.19	15.15	30.33	57.03	100
Brain_Tumor2	83.54	25.28	70.79	86.06	88.67
Leukemia1	57.49	45.29	35.20	89.8	96.9
Leukemia2	97.07	39.07	73.61	100	97.4
SRBCT	95.92	83.03	86.58	92.94	98.82
Prostate_Tumor	87.66	56.72	82.55	78.4	82.6

Table 4 # of Genes Selected

Dataset	SVM	BPNN	CART	PSODT	NBBFA
Lung cancer	12601	12601	12601	5213	4946
Brain_Tumor1	5921	5921	5921	4902	2203
Brain_Tumor2	10368	10368	10368	4171	3983
Leukemia1	5328	5328	5328	2681	2069
Leukemia2	11226	11226	11226	5578	4384
SRBCT	2309	2309	2309	1095	847
Prostate_Tumor	10510	10510	10510	4064	3853

From Table 4, it is found that NBBFA identified the minimal number of genes than PSODT.

V CONCLUSION AND FUTURE WORK

This study presented an application of the Firefly Algorithm, which contributes to identifying relevant genes causing cancer. In this work, the Naive Bayes classifier guides the binary firefly algorithm for gene selection by evaluating the fitness of fireflies. The fitness parameter is accuracy and result obtained is number of genes selected from the large dataset. Experimental results clearly illustrate that minimum genes are selected using NBBFA compared with PSODT. Also, NBBFA has improved classification accuracy when compared with PSODT. Other nature-inspired algorithms can be used for improving accuracy. Instead of using naive Bayes classifier different classifiers can be used to improve classification accuracy.

REFERENCES

1. Alba E, et al. (2007) 'Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms,' IEEE C Evol Computat, Vol. 9, pp. 284–290.
2. Ben-Dor A, Bruhn L, Friedman N et al (2000) 'Tissue classification with gene expression profiles', J Comput Biol, Vol. 7, pp. 559–583.
3. Chaung et al (2008) 'Improved binary PSO for feature selection using gene expression data' , Computational Biology and Chemistry, Vol. 32, pp. 29–37.
4. Fayyad, Usama, Piatetsky-Shapiro, Gregory Smyth, Padhraic (1996) 'From Data Mining to Knowledge Discovery in Databases'.
5. He W (2004) 'A spline function approach for detecting differentially expressed genes in microarray data analysis' Bioinformatics, Vol. 20, pp. 2954–2963.
6. Kun-Huang Chen, Kung-JengWang, Min-Lung Tsai, Kung-Min Wang, Angelia Melani Adrian, Wei-Chung Cheng, zu-Sen Yang, Nai-Chia Teng, Kuo-Pin Tan9 and Ku-Shang Chang, Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm, Bioinformatics 2017, 15:49
7. Li L, Darden TA, et al (2001) 'Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method', Combinatorial Chemistry and High Throughput Screen, Vol. 4, pp. 727–739.
8. Li S, Wu X, Tan M (2008) 'Gene selection using hybrid particle swarm optimization and genetic algorithm', Soft Comput , Vol. 12, pp. 1039–1048.
9. Mohamad MS, et al (2009) 'Particle swarm optimization for gene selection in classifying cancer classes', Proceedings of the 14th International Symposium on Artificial Life and Robotics, pp. 762–765.
10. Prasad Y, K. K. Biswas, and M. Hanmandlu, "A recursive PSO scheme for gene selection in microarray data," Appl. Soft Comput. J., vol. 71, pp. 213–225, 2018.
11. Ranjani Rani R and D. Ramyachitra, "Microarray cancer gene feature selection using spider monkey optimization algorithm and cancer classification using SVM," Procedia Comput. Sci., vol. 143, pp. 108–116, 2018.
12. Rana F. Najeeb and Ban N. Dhannoon (2018), A feature selection approach using Binary Firefly Algorithm for network intrusion detection system, Vol. 13, No. 6.
13. Sharma A., and Rani R., "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," Comput. Methods Programs Biomed., vol. 178, pp. 219–235, 2019.
14. Shen et al (2008) Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, Computational Biology and Chemistry, Vol. 32, pp. 53–60.
15. Thomas, J.G., Olson, J.M., Tapscott, S.J., Zhao, L.P., 2001. An efficient androbust statistical modeling approach to discover differentially expressedgenes using genomic expression profiles. Genome Res. 11, 1227–1236.
16. Troyanskaya OG, Garber ME, Brown PO et al (2002), 'Nonparametric methods for identifying differentially expressed genes in microarray data', Bioinformatics, Vol. 18, pp. 1454–1461.
17. Wang H, Tan L, and Niu B, "Feature selection for classification of microarray gene expression cancers using Bacterial Colony Optimization with multi-dimensional population," Swarm Evol. Comput., vol. 48, no. April, pp. 172–181, 2019.
18. Xin-She Yang. 2009. Firefly Algorithms for Multimodal Optimization, Stochastic Algorithms: Foundations and Applications. pp. 169-178.
19. Zhao W, et al (2008) 'A novel framework for gene selection', Int J Adv Comput Technol , Vol. 3, pp. 184–191.

AUTHORS PROFILE

Dr. R.R.Rajalaxmi completed M.E. Computer Science and Engineering under Bharathiar University in the year 2000. She finished her PhD in Anna University during 2011. Presently she is working as the Professor and Head of the Department of Computer Science Engineering, Kongu Engineering College, Tamil Nadu. She is life member of CSI and ISTE. Her research areas are Data Mining, Big Data Analytics, Nature Inspired Computing, Machine Learning, and Deep Learning. She organized 9 sponsored seminars/workshops/training prorammes. She published 11 research papers in international journals and 6 articles in international conferences. She also completed 3 research projects sponsored by various funding agencies.

Dr.E.Gothai completed Ph.D in the area of Big Data Analytics. She has 19 years of teaching experience. She published 9 research papers in international journals and presented 16 articles in the conferences. She is working as an Associate Professor in the Department of Computer Science Engineering, Kongu Engineering College, Tamil Nadu. Her areas of research are Data mining, Big Data Analytics, and Deep Learning.

Dr. R.Thamilselvan completed Ph.D from Anna University in the area of Grid Computing. He has 19 years of teaching experience. He published 14 research papers in International journals and presented 17 articles in the conferences. He is working as an Associate Professor in Information Technology, Kongu Engineering College affiliated to Anna University, Chennai,Tamilnadu. His areas of research are Cloud computing, Grid Computing and Nature Inspired Computing.

Ms. P.Gokila Brindha completed M.Sc (Information Technology) from Anna University in the year 2007. She has 11 years of teaching experience. She published 2 research papers in international journals and presented 2 papers in the conferences. She is working as an Assistant Professor in Computer Technology-UG, Kongu Engineering College affiliated to Anna University, Chennai, Tamilnadu and pursuing PhD under Anna University in the area of Data Mining and Machine Learning.

Dr.P.Natesan completed Ph.D in the area Network Security. He has 19 years of teaching experience. He published 6 papers in international periodicals and presented 9 articles in the conferences. He is working as an Associate Professor in Computer Science Engineering, Kongu Engineering College, Tamil Nadu. His areas of interest include Cloud computing, Network Security and Deep Learning.