An Effective Statistical Integrative Algorithm (Aeiapp) For Protein Prediction

D. Narmadha, A. Pravin

Abstract: ***The task of predicting target proteins for new drug discovery is typically difficult. Target proteins are biologically most important to control a keen functional process. The recent research of experimental and computational -based approaches has been widely used to predict target proteins using biological networks analysis techniques. Perhaps with available methods and statistical algorithm needs to be modified and should be clearer to tag the main target. Meanwhile identifying wrong protein leads to unwanted molecular interaction and pharmacological activity. In this research work, a novel method to identify essential target proteins using integrative graph coloring algorithm has been proposed. The proposed integrative approach helps to extract essential proteins in protein-protein interaction network (PPI) by analyzing neighborhood of the active target protein. Experimental results reviewed based on protein-protein interaction network for homosapiens showed that AEIAPP based approach shows an improvement in the essential protein identification by assuming the source protein as biologically proven protein. The AEIAPP statistical model has been compared with other state of art approaches on human PPI for various diseases to produce good accurate outcome in faster manner with little memory consumption.***

Keywords: ***Protein-Protein Interaction; Graph coloring; Essential proteins; Drug discovery; Computational methods; Knowledge discovery.***

I. INTRODUCTION

Systems biology is the study of composite molecules involved in a cell such as genes, mRNA and protein using mathematical modeling. The study of protein interactions is important for finding the vital target proteins which are crucial for the reproduction of an organism. Javad Zahiri et. al [1] explored the task of finding essential proteins is important for finding the protein target for diseases, developing new drugs, functionality of proteins. However, in the research analysis done by Sung Min Han et. al [2] insight on the impact of the deletion of such essential proteins on the resulting network has been analyzed. Wenqi Hu et. al [3] claimed that essential proteins are highly vital for identifying targets for treating diseases without any side effects. The method for identifying essential proteins is mainly classified as laboratory based testing and computational based approaches.

Manuscript published on 30 September 2019.
*****Correspondence Author(s)

D. Narmadha*, CSE Department, Sathyabama Institute of Science and Technology, Coimbatore, India.
D. Narmadha, CSE department, Karunya Institute of Technology and Sciences, Coimbatore, India.
Dr. A. Pravin, CSE Department, Sathyabama Institute of Technology and Sciences, Chennai, India,

Although the laboratory based testing conducted by various researchers such as mass spectrometry by Aebersold, R [4], RNA interference by Agrawal, N. et. al [5], single gene knockouts by Zhu, J. et. al [6], and conditional knockouts by Liu, P et. al [7] are effective. But, they are highly imperative to identify the target proteins for diseases due to its computational time, cost and complexity. Moreover, these approaches are susceptible to high fault rates and result in huge false negative outcomes. To overcome the challenges of laboratory based test, researchers gave insight into computational based approaches and knowledge discovery process. The knowledge discovery process studied by Holzinger, A [8] gave a survey of the importance of applying data mining based approaches to extract unknown patterns and useful proteins from the protein interaction network. The dense sub small graphs retrieved from the protein network correspond to the essential protein. In recent times, researchers introduced a sequence of computational based approaches such as degree centrality (DC)[9], eigenvector centrality(EC)[10], edge clustering(EC)[11], closeness centrality(CC)[12], betweenness centrality(BC)[13]. Degree centrality is the simplest measure which has been used widely analyzed in a graph based network to find the essential proteins. This computes the essential proteins by computing the total incoming edges to every protein in the protein interaction network (PPI). The protein with the maximum value is likely to be a part of the essential one. Closeness centrality computes the average of shortest path from one protein to every protein in the protein interaction network (PPI). Between centrality computes the average of the shortest path that goes through a particular protein in the network. Eigen vector centrality assigns more weight to the protein if they are connected to the influential protein in a network. In recent times, computational based approaches have also been introduced to find the vital target proteins by combining topological and biological properties of protein interaction network. Yook, S et. al [14] argued a strong correlation between the structure of protein interaction network and the biological properties of the protein. SWEMODE algorithm [15] identifies dense sub graph by assigning weights for the protein based on the strength of association between the proteins. The proteins with high weights are assigned higher ranks. This focuses only on finding functionally similar proteins. Stelzl et. al [16] introduced a scoring based system to extract high confidence interactions from protein interaction network using GO annotation approach. Though, the integration of biological information produces an improvement in the prediction accuracy, the biological information is not available readily to perform the analysis.

Even though, researchers are striving towards finding essential proteins by using computational approaches such as [17], [18],[19],[20]. It is still challenging to predict the essential proteins for specific diseases as none of the methodologies reviewed in literature are focused towards finding essential proteins for specific diseases. Topological properties of the protein interaction network have been analyzed by the researchers in recent time to find vital target proteins. The characteristic properties of hub nodes [21] give directions to analyze the impact of removing hub nodes from a protein interaction network. Ning, K et. al [22] gave insight into new reverse nearest neighbor centrality measure explore the essential proteins by computing the ratio of edges from outside a cluster of specified protein to the number of edges from within the same cluster. In the recent study done by researchers, a wide classification of graph clustering based approaches have been developed to extract the target proteins by identifying dense regions in a protein interaction network. In the beginning MCL based clustering was introduced to perform a random walk through the network. This approach identifies the essential proteins by performing matrix multiplication and diagonal scaling functionality. Later on, in the progress of graph based clustering MCODE algorithm [23] was applied to protein interaction network to find highly connected proteins by assessing the connectivity among the neighbors of a random chosen seed protein. Furthermore, in the development of graph based clustering Restricted Neighborhood based Search Algorithm was introduced to partition the network into smaller sub graphs. Most recently, advanced hierarchical based clustering algorithm [24] was developed to partition a graph into smaller sub graphs by eliminating the least number of edges starting at a node with high degree. In the literature study, functional information of the proteins also has been incorporated to discover the protein complexes. Methodologies has also been incorporated to extract the essential proteins based on functional information of protein [25],[26],[27]. There is considerable improvement in the dense sub graph identification by assimilating topological and biological properties of the protein.

In this research paper, the essential proteins for diseases are identified using a novel mathematical model. To our understanding it is first time a graph coloring algorithm is used to extract most vital target proteins from protein network.

II. PROPOSED METHODOLOGY:

In this methodology of finding essential proteins from protein-protein interaction network, the dataset is collected from STRINGDB database. The proposed AEIAPP based method comprises of three key steps to extract the essential proteins from protein interaction network. In the first step, a biological proven protein is chosen and assigned a low numbered color. In the next step, we check the neighbors of the currently chosen protein. The neighbor of the biologically chosen protein is colored with the next color which had not been used by any of the adjacent vertices. Finally, this proceeds until all proteins in a protein interaction network are colored with minimum chromatic colors.

1.1 New method AEIAPP

AEIAPP

1: Input: A Graph G (V, E) from PPI network is represented as an adjacency matrix where 1 notates the presence of an edge between proteins and 0 represents an absence of edge between proteins.

2: Output: Top percentage of essential proteins for diseases such as cancer, diabetes, allergy and NIPAH from protein interaction network based on colors.

3: Step1: From the PPI, choose a biologically proven protein and assign a chromatic color 1(red)

4: Step2: For each of the remaining proteins of the PPI, check whether it is safe to color the adjacent proteins of the currently picked target protein.

5: Step3: Perform three condition checks at the currently selected protein. Firstly, there may not be any proteins adjacent to the currently chosen protein. Secondly, there may be adjacent proteins which are already colored. Thirdly, there may be adjacent proteins which are not colored.

6: Step4: The first and second condition need not be considered for coloring.

7: Step5: The third condition can be checked, by using the formula’s (1) and (2). It is used to find the proteins that are adjacent to the target protein and to check whether it is already colored.

$$G[p][j] == 1 \qquad (1)$$

$$C1 == x[j] \qquad (2)$$

8: Step6: For all the set of the proteins in a PPI, compute its color.

9: Step7: Top percentages of essential proteins for diseases are computed by extracting the primary and secondary colors from PPI.

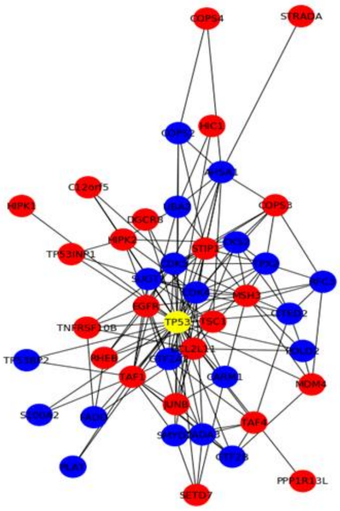

Figure 2: Extraction of primary and secondary interactions from target protein (TP53)

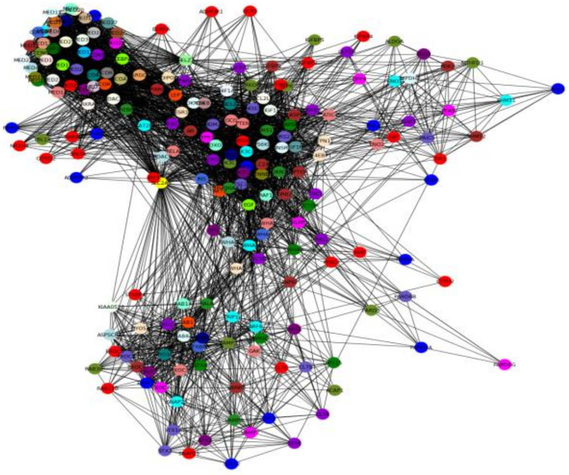

Figure 5: Graph coloring algorithm to protein inetraction dataset

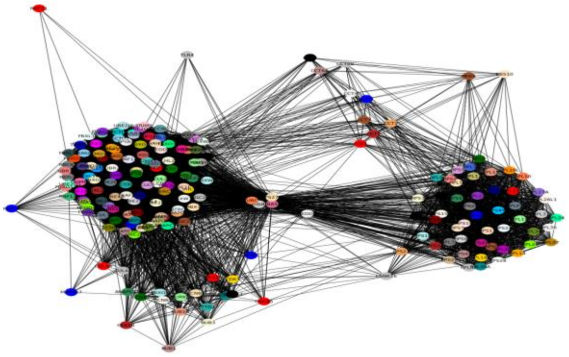

Figure 7: Graph coloring algorithm applied to protein interaction dataset

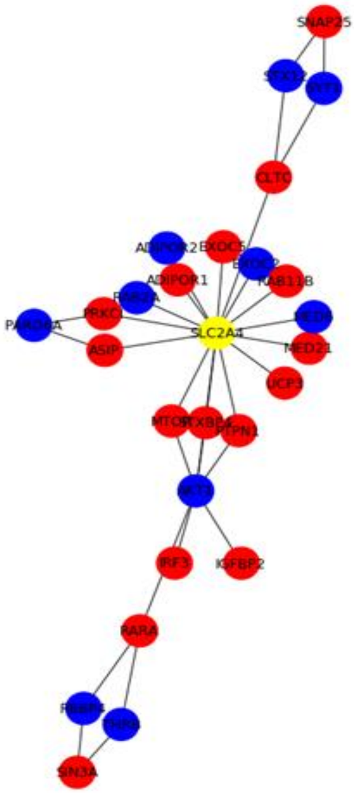

Figure 6: Extraction of primary and secondary interaction from target protein(SCL2A4)

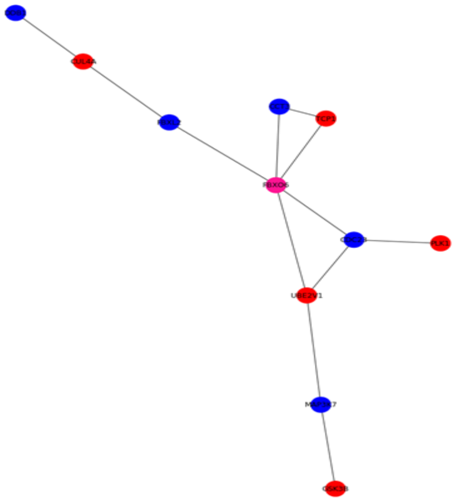

Figure 8: Extraction of primary and secondary interactions from target protein(FBXO6)

III. RESULTS AND DISCUSSION

The experiment was conducted with the huge collection protein interaction data collected from STRINGdb, because the network is complete and reliable. The protein network for different diseases is given as an input to the statistical graph coloring algorithm. The AEIAPP algorithm predicts the relevant target protein for various diseases by extracting the colors. The protein identified with one unique color is categorized as an active target protein. The primary and the secondary colors extracted from the active target protein are categorized as the other relevant essential proteins. The analysis was performed for various human diseases such as cancer, diabetes, allergy and NIPAH virus. The algorithm is able to achieve an average of 85% prediction accuracy.

TABLE I

S.NO	Diseases	No of proteins	No of edges	Target protein	#Essential proteins	Computational Time(sec)	Accuracy(%)
1	Cancer	201	4438	TP53	21	23	89.8
		1587	11348		22	25	87.4
		2587	23689		22	30	86.6
2	Diabetes	257	5331	CDK6	12	32	84.3
		1642	24551		11	31	83.2
		2377	27897		25	31.2	82.5
		3422	36541		22	32.8	81.1
3	Allergy	247	3858	SCL2AC	12	24	79.4
		1374	4724		12	25	80.2
		2174	32794		23	26.5	82.3
		2263	33674		22	29	83.3
4	NIPAH	237	1783		10	32	82.4
		1634	2258		11	33.5	81.6
		2854	32678		12	34	81.5
		3264	33592		13	35	82.4

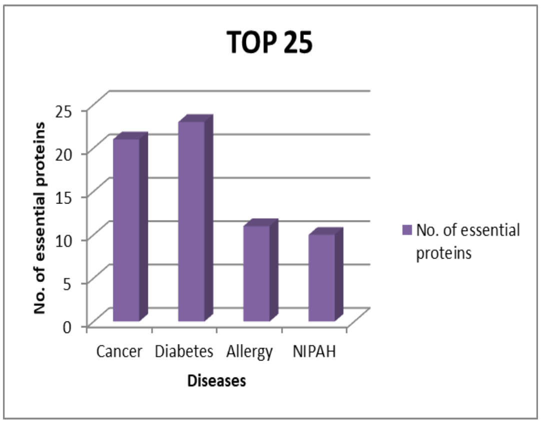

Fig. 9: Top 25% essential proteins

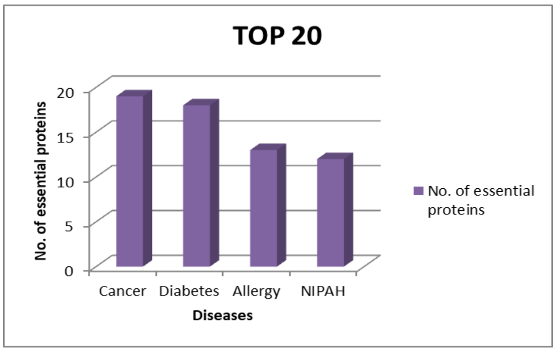

Fig. 10: Top 20% essential proteins

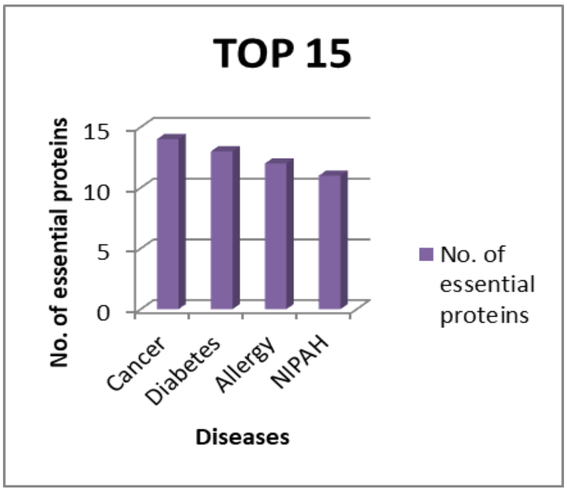

Fig. 11: Top 15% essential proteins

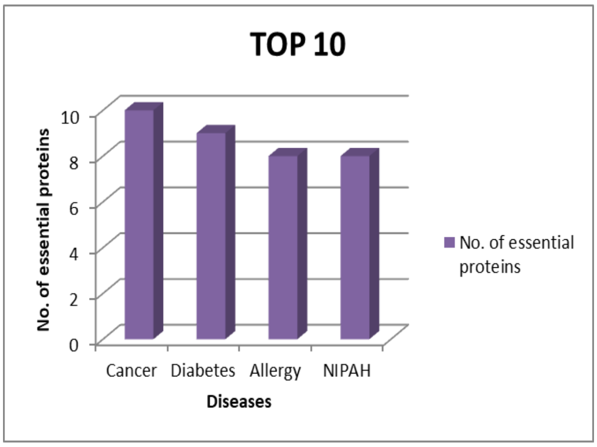

Fig. 12: Top 10% essential proteins

Discussion:

The motivation for extracting the active target proteins from protein-protein interaction (PPI) network is very crucial for developing drugs for diseases, treatment of diseases and to observe the cellular activities. The rapid growth in protein-protein interaction network has posed numerous open challenges for extracting the most essential proteins. The experimental results which have been carried out based on the investigations conducted on protein interaction network have resulted in high false positive and false negative results and hence the results are not appreciable to be considered for further analysis. StringDB is biological software available to analyze both direct as well as indirect association between proteins in order to extract dense sub graph from PPI (Protein interaction Network). It uses algorithms such as MCL and MCODE clustering algorithm for finding dense sub graphs. Firstly, MCL algorithm

IV. CONCLUSION

In this research we have presented a novel method, AEIAPP, for predicting protein interactions that has higher performance than the state of art approaches. It has higher run time magnitude and consumes quite little memory. AEIAPP is very simple to use and makes protein interaction for entire interactomes in short time. The proposed integrative approach helps to extract essential proteins in protein-protein interaction (PPI) network by analyzing neighborhood of the active target protein.

REFERENCES:

1. Javad Zahiri, Joseph Hannon Bozorgmehr and Ali Masoudi-Nejad*, Computational Prediction of Protein–Protein Interaction Networks: Algorithms and Resources, Current Genomics, 2013, 14, 397-414.
2. Sung Min Han, Tae Hoon Lee, Ji Young Mun, Moon Jeong Kim, Ekaterini A. Kritikou, Se-Jin Lee1, Sung Sik Han, Michael O. Hengartner and Hyeon-Sook Koo, Deleted in cancer 1 (DICE1) is an essential protein controlling the topology of the inner mitochondrial membrane in C. elegans, Development 133, 3597-3606 (2006)
3. Wenqi Hu, Susan Sillaots, Sebastien Lemieux, John Davison, Sarah Kauffman, Anouk Breton, Annie Linteau,Chunlin Xin, Joel Bowman, Jeff Becker, Bo Jiang, and Terry Roemer, Essential Gene Identification and Drug Target Prioritization in *Aspergillus fumigatus*, PLoS Pathog. 2007 Mar; 3(3): e24.
4. Aebersold, R. (2003). A mass spectrometric journey into protein and proteome research. *Journal of the American Society for Mass Spectrometry*, *14*(7), 685-695.
5. Agrawal, N., Dasaradhi, P. V. N., Mohmmed, A., Malhotra, P., Bhatnagar, R. K., & Mukherjee, S. K. (2003). RNA interference: biology, mechanism, and applications. *Microbiology and molecular biology reviews*, *67*(4), 657-685.
6. Zhu, J., & Shimizu, K. (2005). Effect of a single-gene knockout on the metabolic regulation in Escherichia coli for D-lactate production under microaerobic condition. *Metabolic engineering*, *7*(2), 104-115.
7. Liu, P., Jenkins, N. A., & Copeland, N. G. (2003). A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome research*, *13*(3), 476-484.
8. Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics*, *15*(6), I1.
9. Zotenko, E., Mestre, J., O'Leary, D. P., & Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology*, *4*(8), e1000140.
10. Ruhnau, B. (2000). Eigenvector-centrality—a node-centrality?. *Social networks*, *22*(4), 357-365.
11. Wang, J., Li, M., Wang, H., & Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *9*(4), 1070-1080.
12. Koschützki, D., & Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, *2*, GRSB-S702.
13. Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.
14. Yook, S. H., Oltvai, Z. N., & Barabási, A. L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, *4*(4), 928-942.
15. Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, *64*(4), 948-959.
16. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., ... & Timm, J. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, *122*(6), 957-968.
17. Skolnick, J., & Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in biotechnology*, *18*(1), 34-39.
18. He, X., & Zhang, J. (2006). Why do hubs tend to be essential in protein networks?. *PLoS genetics*, *2*(6), e88.
19. Li, X., Wu, M., Kwoh, C. K., & Ng, S. K. (2010). Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC genomics*, *11*(1), S3.
20. Massova, I., & Kollman, P. A. (1999). Computational alanine scanning to probe protein− protein interactions: a novel approach to evaluate binding free energies. *Journal of the American Chemical Society*, *121*(36), 8133-8143.
21. Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, *64*(4), 948-959.

22. Ning, K., Ng, H. K., Srihari, S., Leong, H. W., & Nesvizhskii, A. I. (2010). Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. *BMC bioinformatics*, *11*(1), 505.
23. Xue Zhang, Marcio Luis Acencio and Ney Lemke, Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review, Front. Physiol. 7:75. doi: 10.3389/fphys.2016.00075.
24. Gary D Bader and Christopher WV Hogue, An automated method for finding molecular complexes in larger protein interaction network, BMC Bioinformatics, 2003, 4:2.
25. Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, *64*(4), 948-959.
26. Vella, D., Marini, S., Vitali, F., Silvestre, D., Mauri, G., & Bellazzi, R. (2018). MTGO: PPI Network Analysis Via Topological and Functional Module Identification. *Scientific reports*, *8*(1), 5499.
27. Sardiu, M. E., Gilmore, J. M., Groppe, B., Florens, L., & Washburn, M. P. (2017). Identification of topological network modules in perturbed protein interaction networks. *Scientific Reports*, *7*, 43845.

Bibliographic Note:

D.Narmadha received the B.E degree in Computer Science & Engineering from Francis Xavier Engineering College, Anna University, India in 2006, M.Tech degree from the School of Computer Science and Technology, Karunya Institute of Technology and Sciences, Coimbatore, in 2011. She is currently working towards her Ph.D degree in the department of Computer Science and Engineering, Sathyabama University. Her research interests include bio data mining, graph analytics and web mining.

Dr.A.Pravin received the B.E degree in Computer Science & Engineering from Bharath Niketan Engineering College, Madurai Kamaraj University, Madurai, India in 2003, M.E degree in Computer Science & Engineering from Sathyabama University, Chennai, India in 2005 and Ph.D degree in Computer Science & Engineering at Sathyabama University, Chennai, India in 2014. He works currently as an Assistant Professor for the Department of Computer Science and Engineering at Sathyabama University, Chennai and he has 14 Years of teaching experience. He has participated and presented many Research Papers in International and National Conferences and also published many papers in International and National Journals. His area of interests includes Software Engineering, Data mining , Internet of Things and Big data.