International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.070

Identify The Gene Function Using The Multisource Association of Genes By Integration of Clusters Bayesian Network Model

S.D. Lalitha,
Assistant Professor, Department of Computer Science and Engineering,
R.M.K Engineering College,
Kavaraipettai, Tamil Nadu,
India, sdl.cse@rmkec.ac.in

Mohd. Shaikhul Ashraf,
Assistant Professor,
Department of Botany,
HKM Govt. Degree College
Bandipora, Kashmir-193505, mohdshaikhulashraf@gmail.com

R.Bhuvaneswari,
Assistant Professor,
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita VishwaVidyapeetham, Chennai,
India, bhuvanacheran@gmail.com

J.SujiPriya,
Assistant Professor,
Department of Master of Computer Applications,
Sona College of Technology, Salem-636 005,
Tamil.Nadu. sujipriya@sonatech.ac.in

PadmajaNimmagadda,
Professor of Electronics and Communication Engineering,
Mohan Babu University,
(Erstwhile SreeVidyanikethan Engineering College),
Tirupati, Andhra Pradesh,
India padmaja.n@vidyanikethan.edu

Bhasker Pant,
Professor,Department of Computer Science & Engineering,
Graphic Era Deemed to be University,
Dehradun,Uttarakhand,India,
bhasker.pant@geu.ac.in

Abstract—Genomic sequencing is no longer new, the annotation of gene function continues to be a major challenge in biological evolution.From traditional approaches such as affinity precipitation to contemporary broadband approaches such as gene expression microarrays, there are several methods for testing functional genomics.Professional information concerning relative levels of accuracy of data sources is explicitly incorporated into the system for mixing in the normative framework.Multisource Association of Genes by Integration of Clusters (MAGIC) has a level of confidence that enables users to change the rigour of forecasts through Bayesian Model. Researchers used MAGIC to analyze Saccharomyces cerevisiae genetic & physical interactions, micro-array, as well as transcription factor binding site information, but the researchers used Gene Ontology annotations from the Saccharomyces Genome Database to determine a biological value for gene clusters.Compared to microarray analysis, MAGIC simply improves the reliability of functional clustering by building functional clusterings based on a variety of data types.

Keywords—Genomics; Genes Integration; Bayesian model; Gene function

I. Introduction

Increasing levels of high throughput biological information have become accessible in recent times. Some of these, including protein to protein interactions research, affinity precipitation, 2 hybrid approaches, synthetic rescues & lethality tests, as well as microarray analysis, evaluate functional interactions b/w gene products on a wide scale[1]. Although many proteins' functions are unknown, in even model organisms, high throughput information might well be crucial for assigning correct functional annotation on a broad scale [2]. Experimental research could benefit from such predictions since they provide particular hypotheses of evaluation. Some high throughput approaches, on the other hand, forgo selectivity in favor of scalability. By examining co-expression associations in a high-throughput manner, micro-array research could yield gene function estimates. Micro-array data alone frequently lacks the level of sensitivity required for good gene function predictions, however, gene co-expression information is a useful tool for hypothesis creation. An improvement in precision is required for such reasons, even if it comes at the expense of certain sensibility [3]. The integration of diverse functional information inside an integrated analysis could achieve such an increased insensitivity.

Another element of GO is GO annotation, which contains the previously available functional data of genetic variants. Every positive annotation connects a gene to a GO term, indicating that the gene's product performs the function specified by phrase [4]. Every negative annotation, on either hand, implies that the gene product doesn't function properly stated by its word. A Go collaboration annotates genes with GO keywords using model organisms of broad interest to biologists, including Homo sapiens, individually or collectively [5]. Although, our understanding of the functional taxonomy of gene products is still very much in infancy. As a result, both GO annotations & hierarchy & were updated on the regular basis with fresh information & saved for future reference [6]. A Go collaboration annotates genes with GO keywords using model organisms of broad interest to biologists, including Homo sapiens, individually or collectively [5]. Although, our understanding of the functional taxonomy of gene products is still very much in infancy. As a result, both GO annotations & hierarchy & were updated on the regular basis with fresh information & saved for future reference [6].In Dec 2019, GO had grown to over 45,000 words, with every gene annotated only with a few / hundreds of them. As a result, accurately inferring the relationships b/w the genes, as well as the various GO keywords, is hard [8].

Every GO term could be described as a semantic label, making the gene function prediction job a classification task to identify if the label is negative/positive for the gene.

Earlier gene function prediction approaches simply took this annotation data & turned it into a simple binary classification task. As a result of ignoring the relationships b/w GO terms as well as the uneven features of terms, these techniques were inaccurate [9]. Some researchers characterize gene function prediction as just a multiple-label / prediction of structural output problem [10] because a gene is frequently labeled with a group of structurally ordered GO terms at the same time. Some attempted to exploit the inter-relationships between GO keywords & proposed a range of multiple-label learning-based methods.

Various studies that used various sources of data to make functional predictions have demonstrated the utility of merging gene clusters derived from multiple techniques. Other researchers had devised approaches for combining gene expression data including 2 / 2 non-microarray sources of data, resulting in more accurate functional annotation [11]. In gene function predictions, a universal approach of combining diverse high-throughput biological information is required. MAGIC [12] is a probabilistic flexible system fora comprehensive examination of high-throughput biological data information that researchers propose here.

The present aspect of the software is of S. cerevisiae, in which there are numerous interesting data inputs. The technique uses a Bayesian n/w to predict if 2 proteins are functionally connected by combining evidence from various sources of data [13-14]. The n/w effectively conducts a probability "weighting" of sources of data, enabling for structured description of expert knowledge about procedures & avoiding duplicate counting evidence [15]. Each projected functional link is given a posterior belief, which allows users to adjust the prediction's rigor.

In this paper, we introduce MAGIC & show how it may be used to analyze physical & genetic interactions, data on experimentally discovered transcription factor binding locations, as well as a set of data for stress response expression on S. cerevisiae [16-17]. Researchers demonstrate that MAGIC could consistently incorporate non-expression biological information into micro-array analysis, which is impossible to do via simply adding such pairwise complexity information to micro-array grouping systems. Researchers show that, when compared to its input devices, MAGIC improves the accuracy of projected functional connections. Top gene clusters created by MAGIC are described, as well as functional estimates based on it.

II. Materials And Methods

The MAGIC system is designed on a distributed architecture that allows for the addition of new source modalities & sets of data with ease. MAGIC is a general framework that could handle a wide range of data formats & micro-array research techniques. Its n/w incorporates interactions between yeast proteins from the General Repository of Interaction Datasets (GRID)& pairings for the gene from Promoter Dataset for Saccharomyces cerevisiae which had experimentally confirmed binding affinity for the same transcriptional activation. In addition, K means grouping, self-organizing map, & clustering method are all included in MAGIC. A system's inputs were gene clusters based on co-expression or even other empirical studies. A Bayesian network, which is the key part of MAGIC, integrates evidence from input clusters to build a posterior view about whether every gene i to gene j pair does have a functional link. MAGIC effectively poses the subsequent questions for every set of genes: And what's the chance that the product of genes i& j had a functional link, based on the evidence presented.The biological process was defined as a systematic collection of molecular processes aimed at achieving a certain biological goal, such as metabolism. The concept is based on Gene Ontology (GO) Consortium's description for biological mechanisms.

A Bayesian network takes as input gene to gene connection matrix, every reflecting one source of data, with component si,j 0 indicating whether genes I & j were thought to get a relationship of functional &si, j 0 indicating when they're not. Because every matrix is created using a different approach, the criteria of functional relationships for every input sequence are determined by the method utilized to produce the matrix.The intensity of every technique's belief in the relationship that exists b/w genes I & j is represented by the score si,j. This score might be a discrete or binary continuous variable. Because the flexible input format permits genes to belong to several groups/clusters, bi-grouping & fuzzy grouping approaches were not excluded.

An output format was identical to that of the source. MAGIC could accommodate any sort of gene to gene clusters, including protein to protein interactions information, grouping technique outputs, as well as sequence-based information, thanks to the versatility of its I/O formats. MAGIC is written in C++ as well as runs on Linux, with a web-based interface in the works.

III. Network Structure

Researchers engaged specialists in micro-array analysis as well as yeast molecular biology to build a Bayesian network model that effectively depicts links b/w evidence from diverse data sources for purposes of ensemble analysis while avoiding double-counting of data. The structure, as illustrated in Fig.1, integrates sources depending on the type of link found. It makes several independent constraints to enable the much highcorrect populating fordepending upon yeast specialists conditional tables input. These independent assumptions were unlikely to alter the results, considering the sparse environment of non-microarray experimental information. Furthermore, because the techniques represented inside the n/w have diverse underlying concepts, its integration for functional analysis is robust. Specialists of seven in the area of yeast molecular genetics formally examined the previous probability. When the experts were interviewed separately, they showed a high level of concordance in his past beliefs. A PATHFINDER N/W for pathology detection, for eg., had effectively utilized the method of generating Bayesian networks probabilistically offered by specialists in the area.

IV. Measurement Procedures

To assess a gene clustering's integrity, researchers must assess the biological relevance/correctness for gene-gene functional couples which belong to that gene cluster. The key criterion in assessing couples of genes having expected functional links is biological significance, although it is a challenging metric to quantify. Is this a relevant grouping, an experimental mistake, or a biological finding When genes i& j were anticipated were get the connectionof functional nonetheless, there is no biological information related to their features? Although there was no ideal gold standard of gene clusters, annotations under curator control for an S. cerevisiae genome over GO keywords reflect present biological knowledge and thus serve as a fair biological basis for evaluating functional pairings of S. cerevisiae genes. (i)molecular activity, (ii) biological process, & (iii) cellular element are the 3 types of terminology in GO. Every gene could be annotated with one / much more GO keywords from different regions of the go tree since GO has a hierarchical system with the inheritance of multiple. If genes with the same GO annotation phrase on the biological ontology of procedure were thought to be involved in the same biological procedure, researchers focus here on the biological process portion of GO for such an assessment. This is the most important aspect of the ontology of evaluating genes cluster depending on the existence of a functional relationship, even though genes with the same GO annotation phrase on the biological ontology of procedure were assumed to be involved at the same mechanism of biological.

The major challenge in generating physiologically relevant gene clusters appears to be specificity, not sensitivities, due to the high expense of follow-up experimental inquiry. Unfortunately, estimating specificity and sensitivity in S. cerevisiae necessitates information of the entire amount for true positives (TP) &TN couples forgenes that are connected, which is presently hard to estimate effectively. As a result, a percentage of TP pairings is used to evaluate the accuracy of every approach. In its projections, percentage TP method = number of couples predicted by a technique those have a GO term allocation in common/total number for couples predicted via technique, in which TP pairings were also described as couples of genes i& j which have an over-lapping GO annotation of the term:

Gene to gene connection matrices containing gene clusters provide the anticipated couples for every input function, as stated previously. MAGIC systematically includes multiple gene clusters, producing posterior probability of functional relationships b/w every set of genes in the yeast genome.Researchers may compare MAGIC's performance in various ranges of stringency to its input devices since the stringency for MAGIC's projections could be changed by setting the cutoff of the posterior views needed to deem 2 genes operationally linked. By adjusting the cutoff for a score, a median correlation of 2 genes (A, B) and to group's central point (c) that were each part for researchers may vary the stringency of source clustering algorithms. Whenever these clustering approaches were employed commonly for micro-array analysis, really no process is done. Researchers circumvent the problem of favoring techniques that forecast a lower amount of couples in this assessment by evaluating the efficiency of the source clusteringalgorithm as well as MAGIC at every stringency factor.

Researchers utilize MAGIC will be used to connect protein-protein interactions from S. cerevisiae& transcription factor binding sites information with grouping analyzes of a stress response micro-array set of data to demonstrate the potential of MAGIC for the integrative framework of diverse biological data. Researchers evaluate the accuracy for MAGIC's projected functional couples to that of source clustering techniques, demonstrating the effectiveness of MAGIC in merging heterogeneous data.

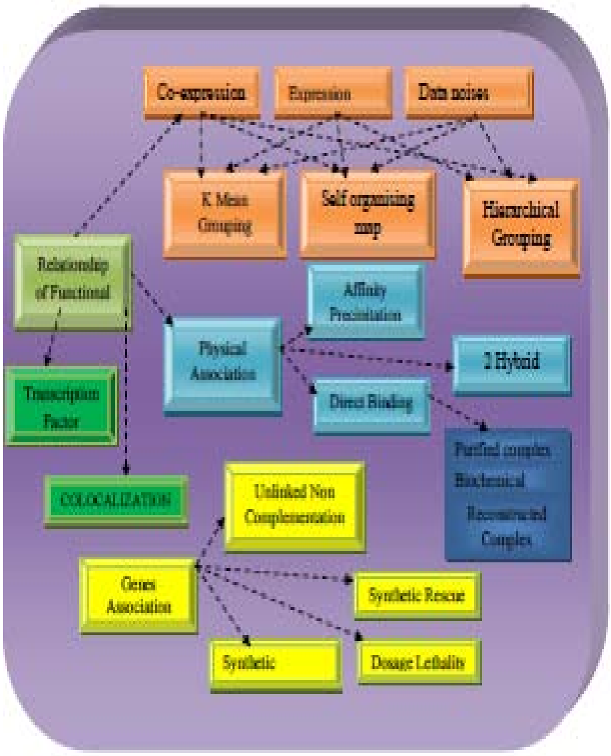

Fig.1.AGIC Bayesian Network

$$S_{A,B} = {}^{x}\!/_{y} \times \int_{k=A,B}^{.} \frac{Cov(k,centroid\ _{c})}{\emptyset_k \emptyset_{centroind\ \ c}}$$

V. Results and Discussion

By utilizing GO with a gold standard, we can assess the biological importance of genes categories. This method of evaluation wasn't without flaws: GO might contain errors in annotation, as well as the activities of several genes inside the yeast genome, were unknown. The false positive (FP) combination of genes might reflect an actual error or a fresh finding, according to the assessment. But certain prejudices may exist inside a subgroup of the gene that was not presently labeled over GO keywords, There's no reason to think otherwise that all those biases might impair grouping techniques in any way. As a result, this strategy provides a sensible & biologically sound foundation for comparing gene categories. Once compared to its input devices, MAGIC continuously increases the proportion of TP couples by incorporating gene clusters depending on micro-array analysis with the already more reliable non-expression based sources of data, as well as MAGIC continuously increases the proportion of TP couples by incorporating gene clusters based on micro-array analysis with the already more accurate non-expression-based sources of data (Fig.2A). High specificity was essential for constructing biologically relevant gene clusters in functional genomic prediction. As a result, researchers concentrate on the greatest specificity area, in which each technique predicts 1,000 or fewer TP couples (see Fig.2B).

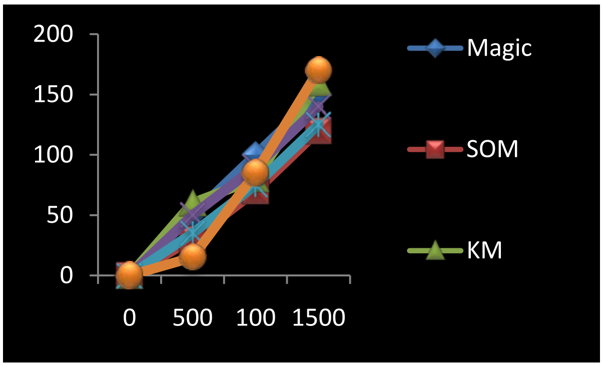

Fig.2A. Non-expression-based sources of data

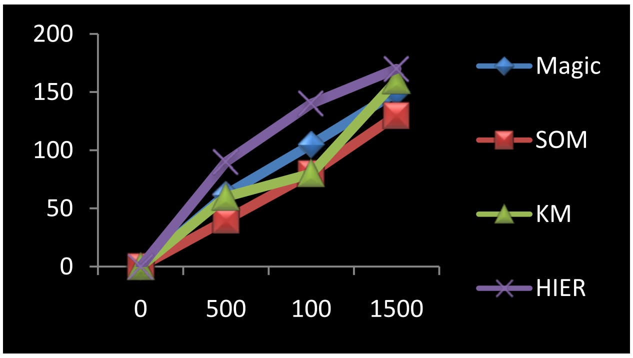

Fig.2B.TP amd FP pairs

When researchers look at the estimates that have the largest percentage of TP couples, MAGIC, which employs non-optimized sources, outperforms the optimized clustering systems, over such a17 percent highlyon the percentage for TP pairings over best forinputssystems but also a most TP couples projected. At very large numbers of projected couples, in which the fraction of TP rates with all approaches are around or below 50percentage, but also at levels unsuitable for precise functional genomic prediction, this variation in efficiency diminishes. As a result, MAGIC generates more physiologically gene that is relevant clusters, over the biggest enhancement inside a highspecificity area.

Micro-array & non-expression information are used by MAGIC. Thus, considering MAGIC's effectiveness based solely on micro-array information / solely on non-expression information is intriguing. MAGIC's accuracy stems partly from the inclusion of non-expression experimental information in the analysis. That's not surprising, then, as MAGIC's efficiency without micro-array information is comparable to the whole system's for a range of over 6,200 projected TP couples (see Fig.2A). A MAGIC application depends on entirely non-expression information that doesn't perform and the complete version as the amount of predictions grows, most likely because it approaches the limit of available information via non-expression sources of data.

Only when micro-array information is considered, however, MAGIC outperforms grouping approaches in the area with such a tiny proportion of couples but has significantly lower TP frequencies than the complete version of MAGIC. When dealing with greater numbers of couples, the micro-array-only system works similarly to the grouping algorithms. As a result, MAGIC utilizes all types of information. It generates extremely precise gene classifications depending on non-microarray experimental information from a multitude of inputs. These clusters were loaded over genes that function in unclear &those functional hypotheses could be generated using micro-array information & other high sensitivity approaches.

Researchers create gene classifications depending on MAGIC's pairwise data by clustering together those genes that have a functional link to the same gene. MAGIC finds groups that reflect the total stress reaction to the surroundings. Such groups are much more specialized for a certain biological mechanism than groups depending on hierarchical grouping which are manually selected. In comparison to a heterogeneous carbohydrate metabolism group depending on hierarchical grouping, MAGIC discovers a group of Snf3, Rgt1, & 5 hexose transporters upregulated in reaction to glucose.A regulation Rgt1, which gets signals via Snf3, a glucose-sensing inside the membranes, induces the transporter in reaction to glucose. With coherent activities involving a multitude of genres, including such protein production (see Fig. 3), MAGIC often detects bigger gene clusters. Protein biosynthesis is ascribed to 49 of 58 identified genes inside the protein biosynthesis group. The research estimates that 10 genes having uncertain annotations inside the group are engaged in protein production.

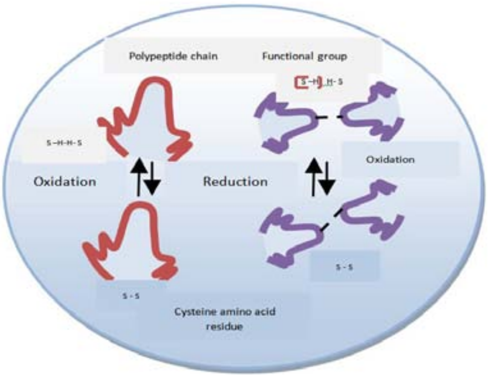

Fig.3. Protein biosynthesis group

As during response to environmental strain, genes responsible for protein breakdown were activated. Refer to Fig.4, MAGIC detects a group of genes associated with ubiquitin-dependent catabolism of protein, gives probable functional annotation of an ORF found in that group, & verifies a previously added annotation of YNL311C. Though Rad23's present GO annotation was "nucleotide-excision repairing, DNA damage detection," it belongs to this cluster. Rad23's role is involved with DNA synthesis probably owing to this prevention for the breakdown repairing protein on DNA damage responding, according to the research. Rad23 connects directly with 26S proteasome & it could play a part in additional degradation of protein processes, according to research. MAGIC categorization indicates Rad23's out-of-date & perhaps deceptive annotation.

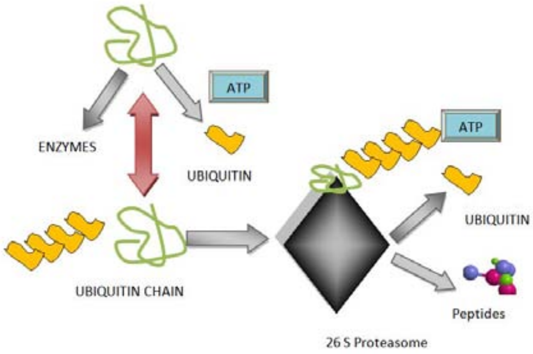

Fig.4.MAGIC detects a group of genes associated with ubiquitin-dependent catabolism of protein

MAGIC thus serves as a durability method of control for current functional classifications for partially characterized genes, in addition to determining the activity of unidentified genes located in clusters well-defined genome. Such examples would be a collection of 3 genes known as PRP8 BUD31& CEF1. PRP8 & CEF1 both are well-splicing enzymes. Depending on a genomic sequence screening of mutations deficient inside the bi-polar budding pattern, BUD31 was currently assigned with bud site location. Although, Snyder & Ni discovered that each number of nuclear proteins, such as those engaged with the processing of RNA, have problems in bud site location, more probably as the result of RNA processing to a gene involved directly with budding. The potential nuclear localization signal had also been discovered in BUD31. Through looking for genes with annotations that don't match some other genes inside a cluster, it's possible to find genes containing erroneous in-complete structural data.

VI. Conclusions

Researchers had demonstrated that MAGIC was another reliable & efficient technique for functional genomic annotation. A method utilizes the probabilistic approach to integrate diverse biological data, resulting in much more physiologically correct gene groups that may be utilized to predict gene function. On combining the outcomes of various algorithms & using the information for biology experts of yeast in the prior probability of a framework of Bayesian, MAGIC avoids the challenge of defining an "ideal" classification approach for microarray data. A system's versatility makes it simple to include different techniques & data sources, and also information from various species.

References

[1] E.H. Mahood, L.H. Kruse, and G.D. Moghe, “Machine learning: A powerful tool for gene function prediction in plants,” Applications in Plant Sciences, vol. 8, no. 7, p. e11376, Jul. 2020.

[2] M. Stamboulian, R.F. Guerrero, M.W. Hahn, and P. Radivojac, “The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction,” Bioinformatics, vol. 36(Supplement_1), pp. i219-26, Jul 1, 2020.

[3] F.P. Obregón, D. Silvera, P. Soto, P. Yankilevich, G. Guerberoff, and R.Cantera,“Gene function prediction in five model eukaryotes based on gene relative location through machine learning,”bioRxiv, Jan 1, 2021.

[4] Y. Zhao, J. Wang, J. Chen, X. Zhang, M. Guo, and G. Yu,“A literature review of gene function prediction by modeling gene ontology,” Frontiers in genetics, vol. 11, p. 400, Apr 24, 2020.

[5] G. Yu, K. Wang, C. Domeniconi, M. Guo, and J. Wang,“Isoform function prediction based on bi-random walks on a heterogeneous network,” Bioinformatics, vol. 36, no. 1, pp. 303-10, Jan 1, 2020.

[6] Y. Cai, J. Wang, and L. Deng,“SDN2GO: an integrated deep learning model for protein function prediction,” Frontiers in bioengineering and biotechnology, vol. 8, p. 391, Apr 29, 2020.

[7] J. Peng, H. Xue, Z. Wei, I. Tuncali, J. Hao, and X. Shang,“Integrating multi-network topology for gene function prediction using deep neural networks,” Briefings in bioinformatics, vol. 22, no. 2, pp. 2096-105, 2021.

[8] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0. Computers and Electrical Engineering, 104, 108481.

[9] T. P. Latchoumi, A.V. Vasanth, B. Bhavya, A.Viswanadapalli, and A. Jayanthiladevi, “QoS parameters for Comparison and Performance Evaluation of Reactive protocols,” In 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), IEEE, pp. 1-4, July, 2020.

[10] D.R. Rani, andG. Geethakumari, “A meta-analysis of cloud forensic frameworks and tools,” In 2015 Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG), IEEE, pp. 294-298, December 2015.

[11] K. Sridharan, andP. Sivakumar, “ESNN-Hybrid Approach Analysis for Text Categorization Using Intuitive Classifiers,” Journal of Computational and Theoretical Nanoscience, vol. 15, no. 3, pp. 811-822, 2018.

[12] K. Sridharan, and P. Sivakumar, “A systematic review on techniques of feature selection and classification for text mining, “International Journal of Business Information Systems, vol. 28, no. 4, pp. 504-518, 2018.

[13] S. Ranjeeth, andT.P. Latchoumi, “Predicting Kids Malnutrition Using Multilayer Perceptron with Stochastic Gradient Descent Predicting Kids Malnutrition Using Multilayer Perceptron with Stochastic Gradient Descent”.

[14] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ...&Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis. Computers and Electrical Engineering, 106, 108556.

[15] C.Bhuvaneshwari, and A.Manjunathan, “Reimbursement of sensor nodes and path optimization”, Materials Today: Proceedings, vol. 45, pp.1547-1551, 2021.

[16] C. Bhuvaneshwari, and A. Manjunathan, “Advanced gesture recognition system using long-term recurrent convolution network”, Materials Today Proceedings, vol. 21, pp.731-733, 2020.

[17] M.Ramkumar, A. Lakshmi, M.P.Rajasekaran, and A.Manjunathan, “MultiscaleLaplacian graph kernel features combined with tree deep convolutional neural network for the detection of ECG arrhythmia”, Biomedical Signal Processing and Control, vol. 76, p. 103639, 2022.