International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

An Effective different Data Mining algorithms for Prediction of Warning level in Aircraft Accident Dataset

Dr.A.B.Arockia Christopher [1], Dr.M.Balakrishnan[2], Dr.A.P.Janani[3]

[1] Assistant Professor (SG) , [2] Assistant Professor (SG) and [3] Assistant Professor (SG)

Department of Information Technology

Dr.Mahalingam College of Engineering and Technology

Pollachi, Coimbatore, Tamilnadu, (India)

ABSTRACT

This paper mainly motivations on different data mining algorithms applied on the huge number of datasets of an airline databases to understand and clean the dataset. The feature selection techniques of Correlation Feature Selector Subset Evaluator, Consistency Subset Evaluator, Gain Ratio feature evaluator, Information Gain Attribute Evaluator, OneR feature evaluator, Principal Components Attribute Transformer (PCA), ReliefF Attribute Evaluator and Symmetrical Uncertainty Attribute Evaluator are used in this analysis in order to reduce the dataset attributes. Also the Decision Tree classifier techniques of data mining are used to predict the warning level of the component as the class attribute in aircraft accidents for Risk and Safety. For this intention Weka software tools are used. This study also demonstrated that the Principal Components Attribute Transformer would performance of accuracy 99.8 percentage better than other attribute evaluators on airline dataset. This work may be useful for Aviation Company to make better prediction.

Keywords — Risk, safety, CfsSubsetEval, PCA, GainRatio.

I. INTRODUCTION

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series. In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many *redundant* or *irrelevant* features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques provide three main benefits when constructing predictive models: improved model interpretability, shorter training times and enhanced generalization by reducing over fitting. Feature selection is also useful as part of the data analysis process, as shows which features are important for prediction, and how these features are related.

Data Mining is an analytic process which designed to explore data usually large amounts of data. We have used different feature selection techniques for discovering attribute Evaluator and generating a decision tree. These techniques are CFS Subset Evaluator, Consistency Subset Evaluator, Gain Ratio feature evaluator, Information Gain Attribute Evaluator, OneR feature evaluator, Principal Components Attribute Transformer, ReliefF Attribute Evaluator and Symmetrical Uncertainty Attribute. Data mining tool Weka is used in this article research paper.

This is particularly so under current conditions of continuous growth in air transport demand, frequent scarcity of airport and infrastructure capacity, and thus permanent and increased pressure on the system components. Risk and safety have always been essential considerations in aviation. With the fast growth in air travel, flight delays, cancellations and incidents/accidents have also dramatically increased in recent years. Aviation accidents may result in human injury or even death. An airline company collects several case reports including structural and textual data.

In this study, we applied different data mining approaches on the incident reports. We, the decision trees and Feature selection algorithms to find the performance of the accuracy about the incidents resulted in fatality. The decision tree techniques of data mining is use to predict the warning level of the component as the class attribute. We have explored the use of the decision tree techniques on aviation components data. We gave PCA rules that were found by this analysis for the experts of Aviation Company. Some safety recommendations are address to the Airline Aviation Administration.

The paper is organized as follows. In Sections 2 and 3 we mentioned the related work and background of feature selection. In Section 4 we described the dataset descriptions about classification rule discovery for the aviation incidents resulted fatality.

II. RELATED WORK

The area of data mining and knowledge discovery is inherently associated with databases. Data mining methods are used in the process of knowledge discovery to reveal new pieces of knowledge from large databases. One of the stages in that process is a feature selection. A feature selection is usually meant as a process of finding a subset of features from the original set of features forming patterns in a given data set, optimal according to the defined goal and criterion of feature selection. Within this frame, the decision tree classification provides a rapid and effective method of categorizing datasets. For many problems of classification where large datasets are used and the information contained is complex and may contain errors, decision trees provide a useful solution. Aitkenhead presented a method which combines the decision tree paradigm with different evolutionary concepts to produce a classification methodology.

The US Mine Safety and Heath Administration (MSHA) developed a mine accident database from Part 50 of the Federal mine safety regulations. This database has been used to track the numbers, rates and severity of mine accidents in the United States. Epidemiologists and mine safety researchers have still used it to perform many analyses, helping to guide research and best practice. Dessureault and his friends explored the background of the Part 50 database, give a general background of data warehousing and data mining, and present some of the interesting analyses that resulted from a modernized Part 50 data warehouse using data mining. Shyur [13]

proposed a model that allows investigation of non linear effects of aviation safety factors and flexible assessment of aviation risk.

The main goal of the research is to develop a model to provide relative risk probability inference and trend analysis among different kind of human errors which may cause any major aviation events. A subset of data gathered from the Flight Safety Management Information Systems (FSMIS) developed by the office of the Taiwan Civil Aeronautics Administration (CAA) was applied to accomplish the study.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

Feature selection, as a pre-processing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. The evaluation functions may be used with different purposes inside the feature selection process.

The information gain measure is biased towards tests with many outcomes. That is, it prefers to select attributes having a large number of possible values over attributes with fewer values even though the later is more informative. For example consider an attribute that acts as a unique identifier, such as a student id in a student database. A split on student id would result in a large number of partitions; as each record in the database has a unique value for student id. So the information required to classify database with this partitioning would be $Info_{studnetID}(D) = 0$. Clearly, such a partition is useless for classification.

Information gain (IG) is based on the concept of entropy. The expected value of information gain is the mutual information of target variable (X) and independent variable (A). It is the reduction in entropy of target variable (X) achieved by learning the state of independent variable (A) [2]. The major drawback of using information gain is that it tends to choose attributes with large numbers of distinct values over attributes with fewer values even though the later is more informative. Correlation based feature selection is the base for symmetrical uncertainty (SU). Correlation based feature selection evaluates the merit of a feature in a subset using a hypothesis – "Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other" [25]. Symmetric uncertainty is used to measure the degree of 14 associations between discrete features. Relief was proposed by Kira and Rendell in 1994. Relief is an easy to use, fast and accurate algorithm even with dependent features and noisy data [25]. The algorithm is based on a simple principle. Relief works by measuring the ability of an attribute in separating similar instances.

Data mining applies data analysis and discovery algorithms to perform automatic extraction of information from vast amounts of data. This process bridges many technical areas, including databases, human-computer interaction, statistical analysis, and machine learning. A typical data-mining task is to predict an unknown value of some attribute of a new instance when the values of the other attributes of the new instance are known and a

collection of instances with known values of all the attributes is given. In many applications, data, which is the subject of analysis and processing in data mining, is multidimensional, and presented by a number of features. Hence, the dimensionality of the feature space is often reduced before classification is undertaken. Feature extraction (FE) is one of the dimensionality reduction techniques. FE extracts a subset of new features from the original feature set by means of some functional mapping keeping as much information in the data as possible. Conventional Principal Component Analysis (PCA) is one of the most commonly used feature extraction techniques. PCA extracts the axes on which the data shows the highest variability. There exist many variations of the PCA that use local and/or non-linear processing to improve dimensionality reduction, though they generally do not use class information.

In our research, beside the PCA, we discuss also two eigenvector-based approaches that use the within- and between-class covariance matrices and thus do take into account the class information. We analyse them with respect to the general task of classification, to the learning algorithm being used and to dynamic integration of classifiers (DIC). During the last years data mining has evolved from less sophisticated first-generation techniques to today's cutting-edge ones. Currently there is a growing need for next-generation data mining systems to manage knowledge discovery applications. These systems should be able to discover knowledge by combining several available data exploration techniques, and provide a fully automatic environment, or an application envelope, surrounding this highly sophisticated data mining engine.

III. FEATURE SELECTION

Feature subset selection is of great importance in the field of data mining. The high dimension data makes testing and training of general classification methods difficult. In the present paper number of filters approaches namely Correlation based feature selection, Consistency Subset Evaluator, Gain Ratio, Information Gain, OneR feature selection; Principal Components Attribute, ReliefF Attribute and Symmetrical Uncertainty Attribute have been used to illustrate the significance of feature subset selection for classifying aircraft database. The Ranker Attribute uses gain ratio to determine the splits and to select the most important features. Best First algorithm is used as search method with Correlation based feature selection as subset evaluating mechanism.

3.1. Correlation based feature selection

In this section, we discuss how to evaluate the goodness of features for classification. In general, a feature is *good* if it is *relevant* to the class concept but is not *redundant* to any of the other relevant features. If we adopt the correlation between two variables as a goodness measure, the above definition becomes that a feature is good if it is highly correlated to the class but not highly correlated to any of the other features. In other words, if the correlation between a feature and the class is high enough to make it relevant to (or predictive of) the class and the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other relevant features, it will be regarded as a good feature for the classification task.

3.2. Consistency Subset

The idea behind these measures is that, in order to predict the concept or class value of its instances, a data set with the selected features alone must be consistent. That is, no two instances may have the same values on all predicting features if they have a different concept value. Therefore, the goal is equivalent to select those

features that better allow defining consistent logical hypothesis about the training data set. As the higher the number of features, the more consistent hypothesis that can be defined, the requisite, of a data set having consistency, is usually accompanied with the criterion of finding a small feature set. In any case, the search for small feature sets is the common goal of feature selection methods, so this is not a particularity of consistency based methods.

3.3. Information Gain

A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. The information gain measure is used to select the test attribute at each node of the decision tree. The information gain measure prefers to select attributes having a large number of values. The basic decision tree induction algorithm J48 was enhanced by C4.5. C4.5 a successor of J48 uses an extension of information gain known as gain ratio, which attempts to overcome this bias. The WEKA [8] classifier package has its own version of C4.5 known as J4.8. We have used J48 to identify the significant attributes.

3.4. Principal Component Attribute selection

Principal component analysis (PCA) is a standard technique used to handle linear dependence among variables. A PCA of a set of *m* variables generates *m* new variables (the principal components), *PC1…PCm*. Each component is obtained by linear combination of the original variables [12], that is:

$$PCi = \sum_{j=1}^{m} bi,j \,.\, Xj$$

$$PC=B^{T}\overrightarrow{X}$$

Where *Xj* is the *jth* original variable, *bi,j* the linear factor. The coefficients for *PCi* are chosen so as to make its variance as large as possible. Mathematically, the variation of the original *m* variables is expressed by the covariance matrix. The transformation matrix *B*, containing the *bi,j* coefficients, corresponds to the covariance eigenvector matrix.

3.5. Relief feature selection

It is efficient, aware of the contextual information, and can correctly estimate the quality of features in problems with strong dependencies between features. The key idea of the original RELIEF algorithm is to estimate the quality of features according to how well their values distinguish between instances that are near to each other.

V. DATASET DESCRIPTIONS

For this analysis, we have used the data from the database of a big airline and aircraft aviation. The application is done on one thousand and five hundred data sets to compare the results. The sample attribute data is given in a report format with following categories shown in Table I. As is apparent from Table I, component reports have 181 attributes. The aim of the analysis is to find the attributes that affect the warning levels and a new formulation about it.

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 11, November 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Table I

Description of Sample Datasets Used In Application

Parameter name	Description
Abind	The challenging for the pilot to diagnose in flight
Aflalo	The side of the runway after landing long
Airatt	The aircraft will be put into a round out attitude shortly before it would otherwise contact the ground
Airbrot	The Burnt-out aircraft is clustered on the aft section of the flight deck, clear of the fire area
Arspd	The highest airspeed attained by an aircraft of a particular class
Aliruy	The Aligning with the runway on takeoff should be a no brainer
Altplm	The standard nominal altitude of an aircraft, in hundreds of feet
Apnopre	The pilot-interpreted make use of ground beacons and aircraft equipment such as VOR, NDB
Arhorfa	The view of heading indicator and artificial horizon after an in flight vacuum failure
ATCerr	The service provided by ground-based controllers who direct aircraft on the ground
Autlad	The designed to make landing possible in visibility too poor to permit any form of visual landing
Autpidiso	The autopilot can control the aircraft while the pilot attends to other duties
Autpieng	The autopilot must be turned on using the Autopilot Engage Switch on the far left
Autthrot	The pilot to control the power setting of an aircraft's engines by specifying a desired flight characteristic
Enfa	Engine failure is probably your worst fear as a pilot
Lftenbrof	The left engine broke off
Rghtenbrof	The right engine broke off
Sersvib	Vibration is bad for any piece of machinery
Wnd	Effect of wind shear on aircraft trajectory

V. EXPERIMENTAL RESULTS

As a part of feature selection step we used eight filter approaches (i) BestFirst search with Correlation based feature selection as subset evaluating mechanism (ii) BestFirst search with Consistency subset as measure to select relevant attributes (iii) Ranking search with gain ratio as measure to select relevant attributes (iv) Ranking search with Information gain as subset evaluating mechanism (v) Ranking search with OneR subset as measure to select relevant attributes (vi) Ranking search with Principal component attribute as calculate to select attributes (vii) Ranking search with ReleifF as performance of select attributes (viii) Ranking search with Symmetric as measure to select attributes from Aircraft accidents/incidents Database. Decision tree with 181

attributes gave an accuracy of 99.8% from PCA. The default *K* folds cross validation method with *K* = 10 was used for decision tree.

The Decision tree induction algorithm classifier produced the analysis of the training dataset and the classification rules. In the experiment, the phase of experiment is the evaluation and interpretation of the classification rules using the unseen data. In the experiment we have used one thousand and five hundred instances of the database as a training datasets. The analyses were performed using WEKA environment. This study also proved that the Principal Component Attribute evaluator filters will performance better than other filters on airline data. Please use only datasets which classification rules best for both on Accuracy and on different selection attributes, as shown in Fig. 1.

Table II

Feature Selection Attributes Performance

Search	Evaluator	Selected Attributes	Accuracy
BTF	Correlation feature selector	12	95.4%
BTF	Consistency Attribute	37	90.2%
Ranker	Gain Ratio	90	98.3%
Ranker	Information Gain	90	98.3%
Ranker	OneR Attribute	90	96.2%
Ranker	Principal Component Attribute	67	99.8 %
Ranker	ReliefF Attribute	90	91.3%
Ranker	Symmetric Uncertainty	90	98.3%

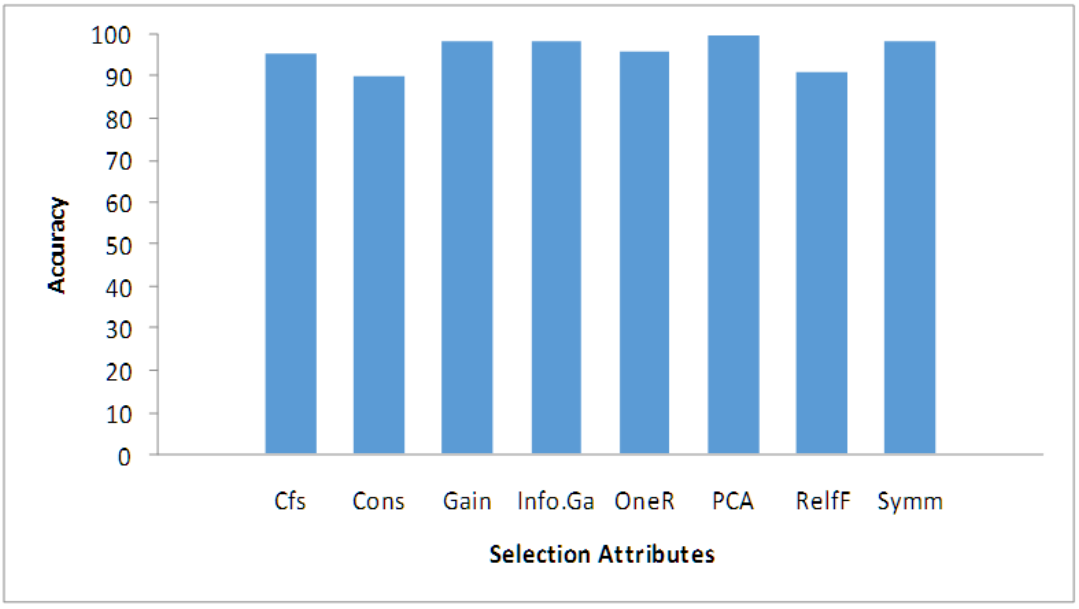

Fig. 1 Performance of different feature selection attributes

VI. CONCLUSION

The feature subset obtained is then tested using classification method namely, Decision Tree Classifier. Experimental results illustrates PCA identified feature subset have improved the classification accuracy when compared to relevant input as identified by decision tree. In this study we have explored the use of different feature selection techniques on aviation accidents data. The main contribution of this study is to evaluate the performance of different feature selection filters are CFS, Consistency, Gain Ratio, Information Gain, Symmetric, Wrapper, ReleifF, OneR and PCA within aviation components data. We talked about the PCA filters that were found by this analysis with the experts of Aviation Company better than other filters. As a result said that PCA filters are some safety recommendations are address to the Airline Aviation Administration.

ACKNOWLEDGMENT

The authors wish to acknowledge and also gratefully acknowledge the unanimous reviewers for their kind suggestions and comments for improving this paper.

REFERENCES

[1] A.B.Arockia Christopher & Appavu alias Balamurugan, S, 'Prediction of warning level in aircraft accidents using data mining techniques', *Aeronautical Journal*, Royal Aeronautical Society, London, ISSN: 0001 9240 in Anna University-Chennai, Annexure I, August 2014, Vol. 118, No. 1206, pp. 935-952.

[2] A.B.Arockia Christopher, V.Shunmughavel and A.B.Antony Anderson, 'Large Scale Data Analysis on Aviation Accident Database using different Data Mining Techniques.', *Aeronautical Journal*, Royal Aeronautical Society, London, ISSN: 0001 9240 in Anna University-Chennai, Annexure I, December 2016, Vol.120, No.1234, PP. 1849-1866.

[3] A.S. Chang, S.S. Leu, Data mining model for identifying project profitability variables, International Journal of Project Management 24 (2006) 199–206.

[4] Asha Gowda Karegowda1, A. S. Manjunath2 & M.A.Jayaram3. " Comparative study of attribute selection using gain ratio and correlation based feature selection", International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 271-277.

[5] C. Apte, S. Weiss, Data mining with decision trees and decision rules, Future Generation Computer Systems (1997).

[6] C.C. Chang, R.S. Chen, Using data mining technology to solve classification problems, A Case Study of Campus Digital Library, Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan, 2006.

[7] Chen, W., Tseng, S., & Hong, T. (2008). An efficient bit-based feature selection method. Expert Systems with Applications, 34, 2858–2869.

[8] D. Hand, H. Manila, P. Smyth, Principles of Data Mining, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England, 2001.

[9] Dunham, M. H. (2002). Data mining introductory and advanced topics part I. Department of Computer Science and Engineering Southern Methodist University.

[10] Feyza Gürbüz, Lale Özbakir, Hüseyin Yapici (2009). Classification rule discovery for the aviation incidents resulted in fatality. Knowledge-Based Systems 22 (2009) 622–632.

[11] Feyza Gürbüz, Lale Özbakir, Hüseyin Yapici (2011). Data mining and pre-processing application on component reports of an airline company in Turkey. Expert Systems with Applications 38 (2011) 6618–6626.

[12] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

[13] H. Jantan et al. *"Classification for Prediction",* International Journal on Computer Science and Engineering, 2(8): 2526-2534, 2010.

[14] H. Jiawei, M. Kamber, Data Mining: Concepts and Techniques, University of Simon Fraser, 2001.

[15] H.J. Shyur, A quantitative model for aviation safety risk assessment, Computers and Industrial Engineering (2007).

[16] J.R. Quinlan, *Induction of Decision Trees*, Machine Learning 1: pp.81-106, Kluwer Academic Publishers, Boston, (1986).

[17] Jiawei, H., & Kamber, M. (2001). Data mining: Concepts and techniques. University Of Simon Fraser.

[18] Kim, Y., Street, W. N., & Menczer, F. (2003). Feature selection. In Data mining. USA: University Of Iowa.

[19] M. Balakrishnan & K. Duraiswamy "Efficient Online Tutoring Using Web Services", International Journal of Computer Science Issues, 2011, Vol. 8, Issue 5, No 3, pp.192-195.

[20] M. Balakrishnan & K. Duraiswamy "Learning Software Component Model for Online Tutoring", Journal of Computer Science, 2012, vol. 8, no.7, pp. 1150-1155.

[21] M. Bineid, J.P. Fielding, Development of a civil aircraft dispatch reliability prediction methodology, Aircraft Engineering and Aerospace Technology 75 (6) (2003) 588–594.

[22] M.J. Aitkenhead, A co-evolving decision tree classification method, Expert Systems with Applications 34 (2006) 18–25.

[23] Mark A. Hall, *Correlation-based Feature Selection for Machine Learning*, Dept of Computer Science, University of Waikato. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.

[24] Nazeri, Z., & Jianping, Z. (2002). Mining aviation data to understand impacts of severe weather on airspace system performance. In Proceedings of the international conference on information technology. IEEE.

[25] Pizzi, N. J., & Pedrycz, W. (2008). Effective classification using feature selection and fuzzy. Integration Fuzzy Sets and Systems.

[26] S. Dessureault, A. Sinuhaji, P. Coleman, Data mining mine safety data, Mining Engineering Littleton 59 (8) (2007) 64. 7 pgs.

[27] S. Solomon, H. Nguyen, J. Liebowitz, W. Agresti, Using data mining to improve traffic safety programs, Industrial Management and Data Systems 106 (5) (2006) 621–643.

[28] S.J. Lee, K. Siau, A Review of Data Mining Techniques, Industrial Management and Data Systems 101/1, MCB University Press, 2001. pp. 41–46.

[29] Shyamala Doraisamy, Shahram Golzari, Noris Mohd. Norowi, Md. Nasir B Sulaiman, Nur Izura Udzir, *A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music* ismir2008.ismir.net/papers/ISMIR2008_256.pdf(2008).

[30] Suraj, Z., & Delimata, P. (2006). Data mining exploration system for feature selection tasks. In International conference on hybrid information technology (Ichıt'06). IEEE, Computer Society. User Manuel of polyanalyst 5, April 2005.

[31] T.C. Hsia, A.J. Shie, L.C. Chen, Course Planning of Extension Education to Meet Market Demand by Using Data Mining Techniques-An Example of Chinkuo Technology University in Taiwan, Taiwan, 2006.

[32] W.S. Tseng, H. Nguyen, J. Liebowitz, W. Agresti, Distractions and motor vehicle accidents: data mining application on fatality analysis reporting system (FARS) data files, Industrial Management and Data Systems 105 (9) (2005) 1188–1205.

[33] www.asias.faa.gov.

[34] www.wikipedia.org

Authors:

Dr.A.B.Arockia Christopher received Ph,D Degree in Information and Communication Engineering from Anna University, Chennai in 2015. He has published more number of books in Lambert Publishing, Germany. He has published more number of journals in Anna University Annexure I and UGC Approved journals. He also published and presented more number of IEEE, Springer and reputed International Conference. He has a recognized supervisor in Anna University, Chennai- 600 025. He is a life time member of Indian Society of Technical Education, New Delhi-110 016.

Dr.M.Balakrishnan received Ph,D Degree in Information and Communication Engineering from Anna University, Chennai in 2014. He has published one book in Lambert Publishing, Germany. He has published more number of journals in Anna University Annexure I and UGC Approved journals. He has a recognized supervisor in Anna University, Chennai- 600 025. He is a life time member of Indian Society of Technical Education, New Delhi-110 016.