International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-5, March 2019

Ample Feature Selection Algorithm for Efficient Prediction of Main Causes of Aviation Accident using Tree based Classifiers

S Sasikala, E A Neeba, A. Suresh, Pethuru Raj, M Hemanth Chakravarthy

Abstract: ***Safety and Happy Journey have always been imperative believe in aviation. Aviation industry has to accumulate huge quantity of experience and data for every year. These data repository includes the data reports including the flight operations, pilot activity report, maintenance report and other supporting reports. Even though these documents are carefully verified for the safest airline journey, it is necessary to provide a precaution checklist with the primary and secondary factor causing the aviation accidents. This paper focus on releasing a Aviation checklist for pre-checking both primary and secondary factors before operating the flight with the help of data mining techniques. The proposed novel feature selection algorithm is compared with traditional feature selection algorithms and its accuracy is evaluated through the Tree based conventional classifiers like J48 (C4.5), Naïve Bayes Tree (NBT), Random Tree (RT), and REP Tree. The research will be justified with real data reports which are collected between the years 1919-2014. This aircraft dataset is provided with 1379 Instances (reports) and 231 attributes (causes). With the classification techniques of data mining, the causes for the aviation accidents are classified as class attribute. The obtained classification accuracy demonstrates that the proposed method could contribute to the successful detection of Aviation Accident Factors and could be applied as pre-check list for the safety journey.***

Keywords: Oscillating Search, Feature Selection, Tree based Classifiers, Correlation based feature selection Aviation Accident Hazards, Improved Oscillated correlation based feature selection.

I. INTRODUCTION

Though air travel is one of the safest methods of transportation, it is expected to double in the next two decades, increasing the aviation accidents risks. Aviation accidents are often shocking incidents that may result in serious injuries or dead. Number of causes of aviation accidents that includes both human and mechanical errors. Apart from these two main causes the other primary causes and secondary causes are reported in this paper for the safety check up to avoid aircraft accidents. The air fact dataset includes the real accident reports from the period 1919-2014. The causes for the accidents were analyzed and based on their occurrence between these periods are considered as primary cause and secondary cause. The majority voted causes are considered as primary cause and rarely occurred cause as the secondary cause. But the severity of both the causes results in the Aviation Accidents.

An air accident is rarely caused by just one event. The figure 1 reveals the fact how data mining techniques fills the gaps in the aviation company.

In this study, the data set has been prepared by considering the real accidental details from the past records and named as "Air Fact" which includes 1379 accident reports as instances and combination of 231 factors with primary or secondary as class or target attributes. Sometimes the Aviation Accident happened due to one reason or sometimes by several factors. Many reports reveal the sole cause for the accidents is pilot's response to an emergency. This time –based reports are now moved towards the feature based reports by using data mining techniques like feature selection and classification. Here a novel feature selector Improved Oscillating Correspondence Dependent Feature Selection developed to validate Air Fact dataset and the features selected by this proposed method is proved as top most causes for Aviation Accidents. This idea reduces the effort to analyses air craft data checklist by filling various Query list. Based on checking the topmost features selected as causes instead of analyzing several factors the check list can be prepared. This enables improvements to be made in aircraft management, affordability, availability, airworthiness and performance to avoid the aircraft accidents. In addition, it highlights the need to evaluate the uprightness of data before take off the flight for the safety air travel.

Typically huge structural set of data could hold complex mistakes in the information. In this case decision tree which is a classification algorithm plays a vital role. This algorithm was proposed by Aitkenhead [1] with the different evolutionary methods. Artificial neural network which is used by classification model is built by Craven et al [2] for binary classification and it has two approaches. Rule extraction and learning the networks are the two approaches. The time consumed is very low. The decision tree which employs the extent of machine learning in which formal rules are mined from a set of observations is proposed by Apte et al [3].

Manuscript published on 30 March 2019.
***Correspondence Author(s)**
S Sasikala, Professor, Department of Computer Science and Engineering, Paavai Engineering College, Namakkal, Tamil Nadu, India.
E A Neeba, Assistant Professor, Department of Information Technology,Rajagiri School of Engineering & Technology, Rajagiri Valley, Kakkanad, Kochi – 682039, Kerala, India.
A.Suresh, Professor & Head, Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, T.M. Palayam, Coimbatore, 641105, TamilNadu, India.
Pethuru Raj, Chief Architect and Vice President, Site Reliability Engineering (SRE) Division Reliance Jio Infocomm. Ltd. (RJIL), SARGOD Imperial, 23, Residency Road, Bangalore 560025, India.
M Hemanth Chakravarthy, Application Development Team Lead, Accenture Technology, Perungalathur, Chennai-63, Tamil Nadu, India.

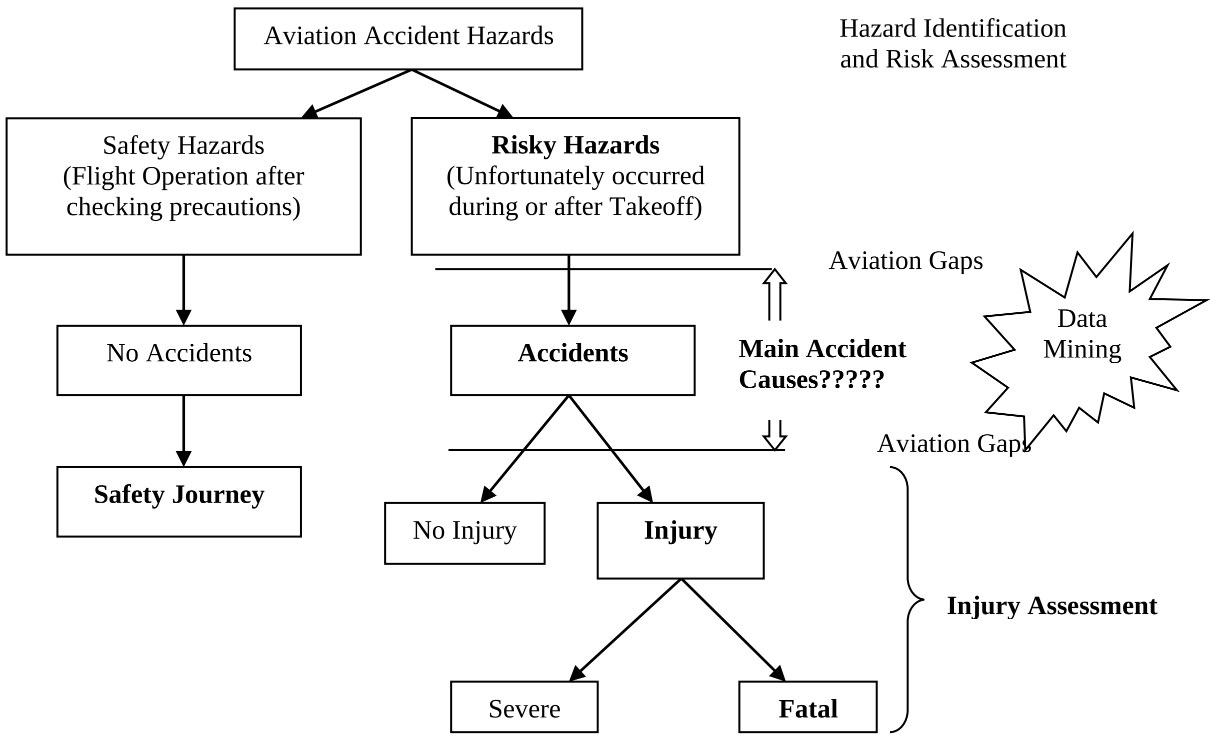

Fig 1: Filling Gaps in Aviation Company using Data mining techniques

Associations circulating the counterfeit financial statements were effectively detected by using data mining techniques is proposed by Kirkos et al. [4]. In aviation industry at National Airspace System, evaluating the climate situation with the help of data mining techniques is done and this was described by Nazeri and Zhang [7]. The data mining techniques are also used in consistency estimation procedure for airplane and it has been made clear in [8].

In [9] introduced replica aimed at interrogating non-linear consequence as aeronautics security factor and flexible appraisal of airplane danger. The relationships flanked by driver alertness and automobile accident are defined in [10].

Numerous accesses for auto-generated choice of topography have been endorsed over years in journalism. After the performance of the feature selection, the results of the classification were better. The other method for feature selection is Correlation Based Filter. Ensemble methods have also gives good result. The goal of this research work is targeted at showing that selection of more significant features from the Air Fact dataset helps the Aviation Company to double check the factors to prevent the Aviation Accidents precautionly. The empirical results show that the proposed IOCFS feature selector achieves remarkable dimensionality reduction in the Air Fact dataset from 231 attributes to topmost features in the order of Top 5, Top 10, Top 15 and so on. The Section 2 here defines the planned procedure by suitable method. Results obtained from the experiment and discussions are described in Section 3 and at last, it is describing with the future scope.

2. Proposed Work- *IOCFS-Improved Oscillated Correlation Feature Selection*

In this work, the information regarding aircraft accident is achieved by two steps: collection of data with preprocessing and classification. The first process associates the collection of data accompanied by feature selection and feature reselection. As aviation accident information are intermediate dimensional, the factors which is the reason for the accidents are not suitably chosen, then the time to figure out the classifier will enlarge. These series of preprocessing steps challenge to minimize the time taken to build the classifier with maximum accuracy. Every individual element accountable for the aviation accident is collected from the earlier information. Such details accumulated in Attribute Relation File Format (.ARFF) and therefore this can be handled by WEKA tool [12]. Few finest reasons from aircraft circumstance data are filtered with Correlation based feature selection (CFS) [13]. As a consequence of this procedure, the feature subset is assembled. Every individual characteristic is weighted with the help of the term correlation frequency. These characteristic which is chosen will be a good demonstrative characteristic of aviation accident dataset and once-again selecting the most excellent by Oscillation search.

The elements are further minimized with the help of an improved search procedure known as "Oscillating search technique(OST)" which was mentioned in our previous work [14]. This search procedure once-again forms a subset with the help of up-down swing method. The finest descriptive attribute in every subset is maintained and the rest are wiped-out.

This search continously customizes the present subset X_d of 'd' features.

This will insert or eliminate feature by using either forward insertion or backward elimination method. With the use of this procedure, we can have the best and eliminate the worst one. The down-swing eliminates at the front and insert at the backside. This complete process forms an oscillation cycle. 'O_{cd}' denotes the depth of the oscillation cycle and this persuades the amount of emphasizes to be chosen in a swing. 'O_{cd}' is enlarged later than the ineffective oscillation cycles and the it is again set to 1 following every individual X_d enhancement. This procedure ends when 'O_{cd}' overtakes a threshold Δ which is defined by the user. Introduction to the special section on managing system change and Mixed-fleet flying in aviation [15][16]. For improving the safety using drilling rig floor [17].

This particular 'd' feature is essential for the oscillatory search. The preliminary set may be accessed accidently or some other methods e.g., with the help of conventional sequential selection measures. The procedure for the Oscillating Search is implemented and is seen in Figure 1. The possible search-restricting parameter is $\Delta \geq \delta$:

1.Start with initial set X_d of features.Set cycle depth to O_{cd} -1.

2.Let $X_{\vec{d}} - ADD^{OST}(REMOVE^{OST}(X_d))$.

3.If $X_{\vec{d}}$ better than X_d ,let $X_d - X_{\vec{d}}$, let O_{cd} -1 and goto 2.

4.Let $X_{\vec{d}} - REMOVE^{OST}(ADD^{OST}(X_d))$.

5.If $X_{\vec{d}}$ better than X_d ,let $X_d - X_{\vec{d}}$, let O_{cd} -1 and goto 2.

6.If $O_{cd} < \Delta$ let O_{cd} - $O_{cd} + 1$ and go to 2.

Figure 2. Pseudo code for Oscillating Search Technique

1.Calculate the Scrore metric of variable set as defined in equation (1).

2.Find the α-the average of Score(S).

3.Rank the Score(S) in descending order.

4.Returning the subsets of α features with the top α weight.

Figure 3. Pseudo code for Correlation based Feature Selection (CFS)

The method of reselection is carried out with CFS executed using Oscillating Search which continuously changes present subset carried out by Correlation based Feature Selection. The restructured feature set is attained by the upswing that includes the improved features obtained to become a fresh subset first which it eliminates poorest and down swings which eliminates the bad features from present subset which inserts improved feature originated for fresh subset. Here, the best and top most features are found as 20 for the 'Ocd'=14.In this particular work, we utilize sequential forward search to exhibit Oscillation Cycle with ADD function which will insert improved one and REMOVE function to eliminate the most awful one.

The feature's selecting processing is shown in figure 3.In figure 3 ,where 'S' is preferred subset of features, means of score metric of the CFS. Therefore, process includes best α feature which is selected and is largely applicable to the classes. The association contribution for most excellent features in a subset should be determined. In fact, the determination of the features is necessary for the aimed classification exhausts. Consequently, the sub-optimal selection method which means selecting again the improved feature among the subset is done with the help Oscillating Search Technique (OST) in figure 4.

1.Let the top selected features α form the feature set by the CFS be SF.

2.For each feature subet apply the Oscillating Search to find best of best features ($X_{\vec{d}}$) until $\Delta \geq \delta$.

3.IF $X_{\vec{d}}$ better than α , then the resulted subset with $X_{\vec{d}}$ features are treated as best features .

4.Return SF′.

4.Validation (v) on SF′ can be calculated as follows :

a.Obtaining the new train data, Tn.X and Tn.Y , in the new feature subset space.

b.Generating a classifier from the training set, using Tn.X and Tn.Y

c.Classifying the validation set data , Valid(X).

$$d.\ v(SF') := \frac{\left|\{x \mid f_S(x) = y, (x, y) \in Validation\right|}{\left|Validation\right|}, x \in \text{Valid}(X), y \in \text{Valid}(Y)$$

e.Return v(SF′)

Figure 4. Pseudo code for proposed IOCFS

II. RESULTS AND DISCUSSION

3.1. Aviation Accident Dataset

This anticipated approach is examined with experimentation which is done on Aircraft Accident dataset. This particular set of data is intermediate dimensional dataset due to its attributes which range in 231 features. The planned procedure creates mutually first-class classification correctness for the Aircraft data. This data is the binary classification dataset which is composed of different classes called primary cause class and ancillary cause class. Amount of examples in the data are 1379 and 231 attributes. The occurrences reported here are actual accident data set which happened at the year 1918-2014.The dataset are accumulated from "Flight Safety Foundation –Accident Prevention" accounts. To demonstrate the efficiency of the anticipated procedure, we have employed a technique - Enhanced Oscillated Correlation Feature Selection (IOCFS) from aircraft area acquired from year 1918-2014. The main source is the actual reasons for the greatest amount of aircraft accident. The inferior reason behind this can be very infrequent. The features of dissimilar set is shown in the Table 1 and Table 2

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 1: Comparison of Proposed Improved Oscillated Correlation Feature Selection (IOCFS) with other Feature Selection Methods

Aviation Accident Dataset	Feature Subset selection with Searching method	CFS with Best First Strategy	Evolutionary Search	Genetic Search	Greedy Step wise	IWSS Embedded NB	Linear Forward Selection	PSO Search	Rank Search	Rerank Search	Scatter Search	Subset Size Forward Selection	Tabu Search	Proposed IOCFS (CFS with Oscillating Search)
(1379, 231,2)	Number of Features Selected	62	55	68	63	64	62	84	82	63	62	63	62	20

Table 2: Performance of Tree based classifier –J48, NB Tree, Random Tree and REP Tree on the topmost causes

IOCFS with Top most causes	J48	NB Tree	Random Tree	REP Tree
Top 200	98.8397	94.8513	98.6222	97.9695
Top 100	98.7672	95.504	98.2596	97.8245
Top 50	98.1871	94.4162	97.9695	97.5344
Top 45	96.3017	94.0537	96.0116	95.7941
Top 40	94.9964	93.1835	94.9239	94.3437
Top 35	94.9964	92.7484	94.9239	94.3437
Top 30	90.7179	91.8782	93.546	93.0384
Top 25	89.1226	91.4431	90.7179	89.9202
Top 20	**98.7672**	**95.504**	**98.2596**	**97.8245**
Top 15	89.2676	91.153	91.008	89.2676
Top 10	89.5577	89.8477	89.7027	88.9775
Top 5	89.05	89.05	89.05	88.905
All	88.9775	91.6606	91.2255	88.5424

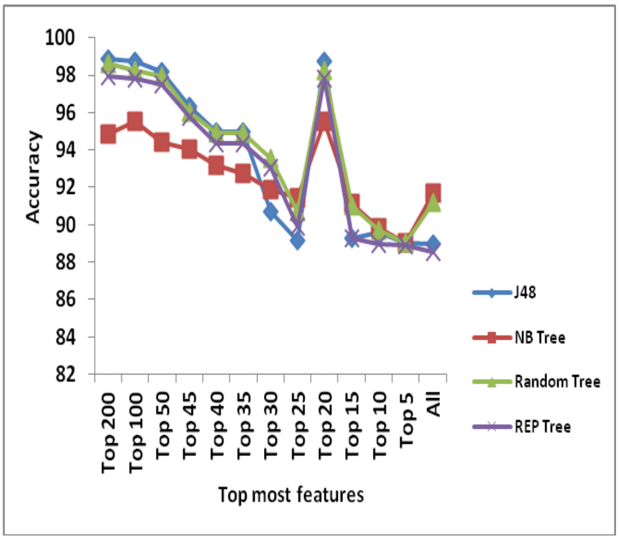

Figure 5.Tree based classifier performance on based Classifier

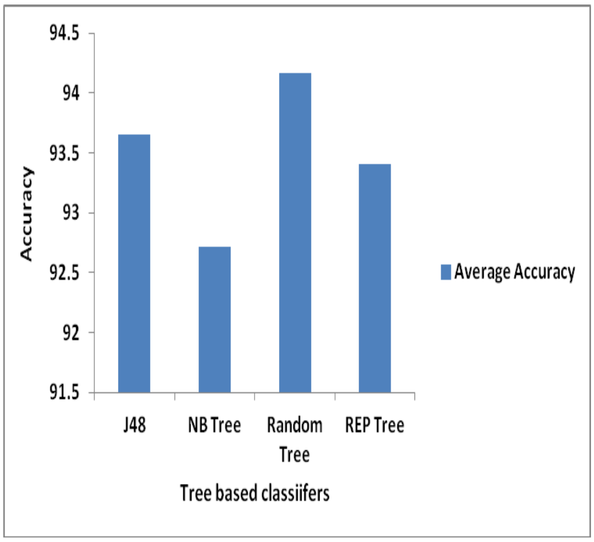

Figure 6.Average Performance of Tree Aviation Data set

3.2. Evaluation of Conventional classifier on Proposed IOCFS Framework

An assessment of planned IOCFS algorithm approved with Conventional Classifiers with help of the features selected by IOCFS. Here the main objective is the amount of features chosen, classifier exactness on the chosen feature subset. Figure 5 reveals the performance of the classifiers, in which J48 results superior performance followed by other classifiers such as NB Tree, Random Tree and REP Tree. Figure 6 shows the random tree as best on the average basis.

III. CONCLUSION

In this article, we focus on the usage of different techniques of mining of data on aviation/aircraft accident data.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

This paper focus on releasing a Aviation checklist for pre-checking both primary and secondary factors before operating the flight with the help of data mining techniques. The proposed novel feature selection algorithm is compared with traditional feature selection algorithms and its accuracy is evaluated through the Tree based conventional classifiers like J48 (C4.5), Naïve Bayes Tree (NBT), Random Tree (RT), and REP Tree. The research will be justified with real data reports which are collected between the years 1919-2014. This aircraft dataset is provided with 1379 Instances (reports) and 231 attributes (causes).With the classification techniques of data mining, the causes for the aviation accidents are classified as class attribute. The essential factor is to concentrate more in the correctness of significant features selected by the proposed IOCFS algorithm.

REFERENCES

1. Aitkenhead,M.J. A co-evolving decision tree classification method, *Expert Systems with Applications*, 34 (1), (2006) 18–25.
2. Craven, M.W. & Shavlik, J.W. Using neural networks for data mining, *Future Generation Computer*
3. *Systems*, 13(1), (1997) 211–229.
4. Apte, C & Weiss,S. Data mining with decision trees and decision rules, *Future Generation Computer Systems*, 1997.
5. Kirkos, E, Spathis. & Manolopoulos, C. Y. Data mining techniques for the detection of fraudulent financial statements, *Expert Systems with Applications* ,32(1),(2007) 995–1003.
6. Nazeri & Jianping Z. Mining aviation data to understand impacts of severe weather on airspace system performance. In Proceedings of the International Conference on Information Technology, (2002) *IEEE*.
7. Bineid, M.& Fielding, J.P.Development of a civil aircraft dispatch reliability prediction methodology, *Aircraft Engineering and Aerospace Technology*, 75(6), (2003) 588–594.
8. Shyur, H.J.A quantitative model for aviation safety risk assessment, *Computers and Industrial Engineering*, (2007).
9. Tseng, W.S.;Nguyen H; Liebowitz, J & Agresti, W.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.
10. Solomon, S.; Nguyen, H, Liebowitz, J.& Agresti, W. Using data mining to improve traffic safety programs, *Industrial Management and Data Systems*, 106 (5), (2006) 621–643.
11. Weka 3: Machine Learning Software in Java. The University of Waikato software documentation. *http://www.cs.waikato.ac.nz/_ml/weka*
12. Devijver, P. A. & Kittler J. *Pattern Recognition*: A Statistical Approach. Prentice Hall, (1982).
13. Somol P. & Pudil P. Oscillating search algorithms for feature selection. In.Proceedings of ICPR 2000, *IEEE* Comp. Soc. (2000) 406–409.
14. Somol, P; Novovicova, P; Grim, J & P. Pudil.Dynamic oscillating search algorithms for feature selection. In Proceedings of ICPR 2008. *IEEE* Comp. Soc. (2008).
15. Chen,Y &Yu,S.Selection of effective features for ECG beat recognition based on nonlinear Correlations. *Artificial Intelligence in Medicine*, 54(1) (2012) 43– 52.
16. corrigan, s. & mcdonald, n. "introduction to the special section on managing system change in aviation: what makes for successful change? cogn tech work (2015) 17: 189. https://doi.org/10.1007/s10111-014-0308-9
17. soo, k., mavin, t.j. & roth, wm. "mixed-fleet flying in commercial aviation: a joint cognitive systems perspective" cogn tech work (2016) 18: 449. https://doi.org/10.1007/s10111-016-0381-3
18. crichton, m.t. "from cockpit to operating theatre to drilling rig floor: five principles for improving safety using simulator-based exercises to enhance team cognition" Cogn Tech Work (2017) 19: 73. https://doi.org/10.1007/s10111-016-0396-9.

AUTHORS PROFILE

Dr. S. Sasikala, currently working as a Professor in Department of Computer Science and Engineering, Paavai Engineering College, Namakkal, TamilNadu, India. She received doctorate in faculty of Information and Communication Engineering from Anna University India, 2016. She has published more than 18 Journal and Conference papers in the area of Data mining and Big Data Analytics with Elsevier Science Direct, Springer and IEEE publishers. She has published two International Scientific books in KDD and Data mining and Data warehousing. She is serving as an Editorial Board Member and Reviewer for many reputed journals like IEEE, ELSEVIER and SPRINGER. She has 16+ years experience in teaching and Research. Prior to joining Paavai Engineering College, she served at K.L.N.College of Information Technology Madurai, Velammal College of Engineering and Technology Madurai, P.S.N.A Engineering college Dindigul and Sethu Institute of technology Madurai. Her research interests include Data Mining, Internet of Things, Machine Learning Paradigms and Optimizations.

Dr. E. A. Neeba, currently working as an Assistant Professor in the Department of Information Technology at Rajagiri School of Engineering & Technology, Kochi, Kerala, which is affiliated to the A.P.J Abdul Kalam Technological University, Kerala. She received her doctoral degree from Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu. She completed her Masters in Computer Science & Engineering from SRM Institute of Science and Technology, Chennai. Her research interests include Analysis of data, Data Mining and Big Data, knowledge representation, and ontology, both from the theoretical perspective and their application to natural language understanding, reasoning, information visualization, and interoperability. Having a rich industrial experience of around 10 years prior to joining academia, and also she has publications in around 10 SCI/ SCIE/Scopus indexed international journals and a few national journals. An active participant in various conferences and workshops on data mining, she is currently involved in several projects in this field. She was entrusted with leadership positions such as the Accreditation coordinator for the college, and Head of the Quality Cell, besides organizing various national and international events.

Dr. A. Suresh B.E., M.Tech., Ph.D works as the Professor & Head, Department of the Computer Science and Engineering in Nehru Institute of Engineering & Technology, Coimbatore, Tamil Nadu, India. He has been nearly two decades of experience in teaching and his areas of specializations are Data Mining, Artificial Intelligence, Image Processing, Multimedia and System Software. He has one patent. He has published 75 papers in International journals. He has published more than 40 papers in National and International Conferences. He has served as a reviewer for Springer, Elsevier, and Inderscience journals. He is a member of ISTE, IACSIT, IAENG, MCSTA, MCSI, and Global Member of Internet Society (ISOC). He has organized several National Workshop, Conferences and Technical Events. He is regularly invited to deliver lectures in various programmes for imparting skills in research methodology to students and research scholars. He has published three books, in the name of Data structures & Algorithms, Computer Programming and Problem Solving and **Python** Programming in DD Publications, Excel Publications and Sri Maruthi Publisher, Chennai, respectively.

Dr. Pethuru Raj has been working as the chief architect in the Site Reliability Engineering (SRE) Center of Excellence, Reliance Infocomm Ltd. (RIL), Bangalore. He previously worked as a cloud infrastructure architect in the IBM Global Cloud Center of Excellence (CoE), IBM India Bangalore for four years. Prior to that, He had a long stint as TOGAF-certified enterprise architecture (EA) consultant in Wipro Consulting Services (WCS) Division. He also worked as a lead architect in the corporate research (CR) division of Robert Bosch, Bangalore.

In total, He have gained more than 17 years of IT industry experience and 8 years of research experience. He obtained his PhD through CSIR-sponsored PhD degree in Anna University, Chennai and continued the UGC-sponsored postdoctoral research in the department of Computer Science and Automation, Indian Institute of Science, Bangalore. Thereafter, He was granted a couple of international research fellowships (JSPS and JST) to work as a research scientist for 3.5 years in two leading Japanese universities. Regarding the publications, He have published more than 30 research papers in peer-reviewed journals such as IEEE, ACM, Springer-Verlag, Inderscience, etc.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

He have authored 7 books thus far and He focus on some of the emerging technologies such as IoT, Cognitive Analytics, Blockchain, Digital Twin, Docker-enabled Containerization, Data Science, Microservices Architecture, etc. He have contributed 25 book chapters thus far for various technology books edited by highly acclaimed and accomplished professors and professionals. The CRC Press, USA had also released his first book titled as "Cloud Enterprise Architecture" in the year 2012 and you can find the book details in the page http://www.crcpress.com/product/isbn/9781466502321 He has edited and authored a book on the title" Cloud Infrastructures for Big Data Analytics" published by IGI International USA in March 2014. A new book on the title" Smarter Cities: the Enabling Technologies and Tools" by CRC Press, USA, is to hit the market in the month of June 2015. He has collaborating with a few authors to publish a book on the title "High-Performance Big Data Analytics" to be published by Springer-Verlag in the year 2015.

Dr M.Hemanth Chakravarthy, currently working as Application Development Lead in Accenture Services. He obtained his M.E (Software Engineering) from GIET, Rajahmundry (JNTU, Kakinada). He Completed his Ph.D in the Department of Computer Science and Engineering at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai-62, Tamilnadu, India. He has been in Software Industry for the past 10 years and has expertise in Salesforce, Software Testing and Java. He has primarily worked in Sales, Services and Marketing Clouds and has expertise in Roles, Profiles, Hierarchies, Workflows, Rules and Validations along with Chatter and Triggers. He has also Extensively worked on integration systems with legacy applications to SFDC His area of research is Cloud Computing. He has published 5 research articles in International Journals and 2 papers presented in international Conferences. He has attended various Training Programmes, Workshops and FDPs related to his area of interest.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)