International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

A Comprehensive Big Data Analytics Based Framework for Premature Recognition of Breast Cancer

Mithra Venkatesan, Anju V. Kulkarni, Radhika Menon, V. Ramakrishnan

Abstract: ***Healthcare industry is fast growing and expanding in rapid pace. The volume and veracity of data generated in the industry is massive and requires huge storages and handling capability. Big data is empowered with such robust abilities and hence most suitable for handing large amount of data. Further, hese data could be utilized towards building predictive and forecasting models. Breast cancer is a deadly form of cancer majorly affecting women around the globe. The concept of big data and predictive analytics is being explored in the paper towards early diagnosis of breast cancer. This paper surveys various literatures available on application of big data analysis for breast cancer. Subsequently a comprehensive framework is being proposed based on the gaps identified. Different machine learning algorithms which can be applied in the framework is also detailed in the paper. Such frameworks when implemented will greatly help in handling the massive data available and aid in early detection of breast cancer.***

Keywords: Big Data, Predictive Analytics, Healthcare, Breast Cancer

I. INTRODUCTION

Healthcare is an information intensive industry. The healthcare industry is changing at a dramatic rate. There are multiple processes going on within the health sector. These processes not only impact the care of individuals but also help medical practitioners and the delivery of care and services. The field of big data and analytics is extremely powerful and is rapidly expanding. These technologies have started playing pivot role in the emerging healthcare industry. The huge volumes of healthcare data can be aggregated and structured using big data based tools. Analytical models based on the collected data could aid in prediction of diseases or in improvisation of health care mechanisms. [1] discusses the major challenges in medical research involving application of big data analytics for different healthcare applications. It also gives a comprehensive overview of different areas of research in the field.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Mithra Venkatesan*, Associate Professor, Dept. Of E&TC, D.Y.Patil Institute of Technology, Pimpri, Pune, India.-18
Anju.V.Kulkarni, Professor, Dept. Of E&TC, D. Y. Patil Institute of Technology, Pimpri, Pune, India.-18
Radhika Menon, Professor, Dept. Of Mathematics, D. Y. Patil Institute of Technology, Pimpri, Pune, India.-18
V.Ramakrishnan, Dept. Of E&TC, Software Architect, Siemens PLM , Pune, India.-18

There are different powerful tools and techniques which could be utilized towards improving existing healthcare facilities [2].The different methods used to store and retrieve in big data analytics when applied to healthcare is discussed in [3]. Further benefits of using analytics techniques are also illustrated. [3]. The big data analytics could be developed based on different comprehensive broad architectures making use of different open sources such as Hadoop, Apache Strom etc. [4][5]. The amalgamation of significant throughput, ability of real time computing and storage capacity can efficiently handle large volumes of data related to healthcare being generated at fast rate.

Internet of Things (IoT) leads to technology of big data and big data based analytics. IoT based healthcare systems and in-hospital systems have been developed towards improving decision support system and predictive diseases [6], [7].The extension of Cloud Computing technology can also be evolved in big data. Hence these two technologies have been successfully merged and being utilized towards finding healthcare solutions to complex problems [8]. In fact over the traditional healthcare techniques, these latest technologies such as big data, data science and analytics actually work towards providing energy efficient low cost health care solutions. [9].

There is growing recognition for Mobile healthcare application with great increase in the availability of sensors in mobile devices. The analytical framework for Mobile health care applications has been discussed [10] along with a case study on the implementation details. Such high volumes of data collected through big data techniques could have huge impacts on the health care industry and the predictive models built through them. This is discussed and detailed in [11].Overall a holistic framework is being built which can act as a reference towards building an analytical ecosystem. [12]

There are different domains of healthcare where big data analytics play a pivot role. Some examples are such as in predicting heart attack and providing medical treatment according to the individual[13], prediction of heart disease [14], prediction of patients with heart failure [15], framework for cardiovascular diseases[16], Chronic Obstructive Pulmonary Disease (COPD) diagnosis in individual patient[17], Cleft-Lip/Palate treatment[18], diabetes management applications[19] etc.

Among different health care domains cancer diagnosis and treatment is a major challenge where cancer is major reason for death of millions of people every year.

There are different types of cancers and each has distinct characteristics and hence diagnosis and treatment changes according to the type of cancer and patient. [20].

Analysis of human genomes can aid in finding problems in DNA and this in turn can find the basic cause of cancer and help in understanding the type of treatment to be giving to the individual. The human genome contains billions of pairs and their analysis is massive. Such massive data sets could be effectively handled by big data. Hence big data analytics is of great help in cancer care, diagnosis, prediction and in treatment. [21]. Just for maintenance of clinical records and for cancer care alone which by itself is a huge task, Big Data analytics could play an important part [22].Other than this, majorly for prediction or diagnosis of cancer analytical models have been built. Work has been done towards building massive data sets for lung cancer [20] and building predictive models towards diagnosis of lung cancer [23].

II. APPLICATION OF BIG DATA ANALYTICS IN BREAST CANCER

There are number of literature available exploring on how big analytics could be used in resolving issues related to breast cancer. Different models have been proposed to tackle breast cancer using various machine learning algorithms.[24] discusses the conceptual prototype which has been developed to identify the presence of breast cancer in early stage using machine learning algorithms. Further, big data is also used in the paper to store all the data which is acquired through learning in the machine learning algorithms. The Bayes Classifier is the algorithm being used in the paper where the software is formulated using python and Wisconsin database is being used. The result got through experimentation is being detailed.[25] explores the usage of Support Vector Machine and Relevance Vector Machine in the same Wisconsin database. Further a new hybrid SVM based model is being developed in the paper and it is found that the hybrid model performs better than the existing SVM and RVM based techniques. Natural Language processing based tools are very efficient in structuring the information present in the health records and accordingly perform analysis on the obtained information from the data. A natural language processing based system has been developed towards processing and structuring clinical reports of patients who are suffering from breast cancer [26].Through this tool it is possible to retrieve various medical diagnosis, characteristics, responses and chances of recurrences based on the analysis done. This method of structuring has good accuracy up to 96%.

In order to perform analytics in big data the tool that is popularly used is Map Reduce. [27] To find the occurrence of breast cancer there is requirement of a tool capable of monitoring and analyzing the end results. Such an efficient tool for big data analytics is MapReduce.This tool not only has the ability to handle large volume of data but can also extract necessary knowledge from this data set and can make decisions based on the knowledge acquired. There is a major requirement to classify the cancer patients in categories of high and low risk groups. This is a kind of predictive model which is to be built. The hybrid combination of Logistic regression and random forests algorithms have been popularly used towards building such predictive models.[28] The initial filtering process is carried out by the random forest technique, while subsequently the interpretable predictive model is built using the Logistic regression methods.

It is well known that the primary method of detection of breast cancer is through mammography. Once cancerous cells are detected it is imperative to find whether they are malignant or benign. Depending on the shape of the mass being malignant or benign could be found using image processing techniques. The finding out of malignant or benign is considered a predictive model and Support Vector Machines(SVM) or back propagation based techniques is being used towards finding them out.[29] The rate of performance, state of training and error histogram are the results depicted based on the predictions. Further, it is indicated that next stage to which cancer belongs can be found out using cloud as well big data based techniques. Towards increasing the classification accuracy and improving the performance, K nearest neighbour algorithms are being employed. [30] The data being provided for the algorithm consists of data being structured, semi-structured as well as unstructured. This data is efficiently being handled by tools available in big data. There are different kinds of data such as clinical data, genomics data, proteomics data which need to be integrated towards building of predictive models. There is requirement towards combining different types of datasets as well as building forecasting models.

Big Data is used towards amalgamation of datasets, Support Vector Machines and Eigen value Decomposition is used for building forecasting models. [31] The proposed model is an efficient method resulting in a mathematical framework for incorporation of data fusion as well as handling non-linear classification issues. There are different mathematical algorithms and classifiers which could be used towards building forecasting models. Comparison of such models enable us make choice over the predictive model to be used for a particular application. The three data mining based algorithms whose ability to predict is compared are Bayes model, Radial Bias Function(RBF) based model and J48.[32] The results point out that Bayes method is superior in terms of high performance accuracy compared to RBF and J48.

There are other techniques available in data mining which are capable building predictive models. Amongst them three techniques namely Decision Tree Support Vector Machine (DT-SVM), Instance based learning and Sequential Minimal Optimization is being compared.[33] It has been found that DT-SVM outperforms the other two methods in terms of prediction accuracy. It is found that data mining algorithms are of tremendous help in prediction algorithms in the early stages of breast cancer. The proposed work in [34] works towards finding the best algorithm that can be used to find the occurrence of breast cancer and improvise the accuracy of the existing algorithms. With the increase risks due to cancer, machine learning based classification methods are popular in diagnosis of breast cancer. Comparison of different machine learning based classification methods are being done to find the beast classifier for breast cancer. [35]

Through literature review, it has been established that machine learning based methods have been very effective in building predictive models for forecasting breast cancer. This paper explores such a model which is capable of taking in the breast cancer big dataset and builds a predictive model capable of forecasting for the same.

III. PROPOSED FRAMEWORK

The three major focus areas for prediction and diagnosis towards prediction of breast cancer are with respect to vulnerability, reappearance as well as survivability. As far as vulnerability is concerned this is with respect to having a complete assessment of the risk before the happening of the breast cancer. The next important spotlight involves reappearance of the disease which handles issues related to redeveloping. Lastly, the third area which is to be highlighted is concerns related to survivability. The various contributing factors for this are related to life expectancy, endurance, development and tumor drug withstanding ability. The accomplishment of any such forecast purely depends on the extent to which the diagnosis is done properly.

The proposed framework developed completely takes into consideration the above three major areas of concerns. Based on the constraints discussed, the high level framework of the proposed method is shown in Fig.1.The basic concept behind the proposed model is that the input is taken as the breast cancer big datasets. These big datasets are subjected to machine learning algorithms. These machine learning algorithms are capable of making forecast or prediction. Based on the forecast the analysis of the performance the predictions are being made.

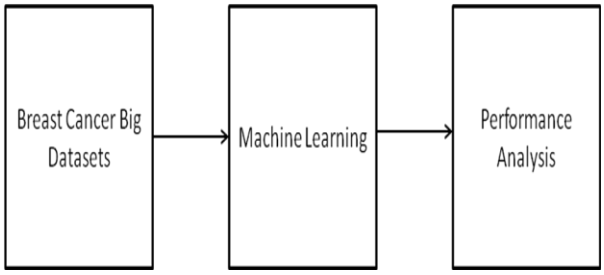

Fig. 1: Block diagram of the process

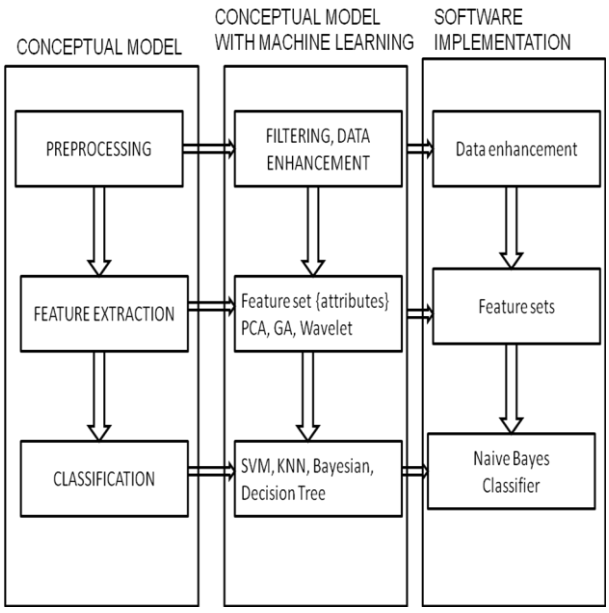

Fig. 2: Abstract model for forecasting of breast cancer based on machine learning algorithms.

The basic steps which are involved in the abstract model are with respect to pre processing, extraction of features and classification. Initially, in the abstract model the pre-processing consists of filtering as well as enhancement of data. Following initial processing being done; the essential characteristics which are required are taken out in order to generate meaningful describing labels from the acquired dataset. This involves the stage of feature extraction wherein the feature is used to classify the cancer in the dataset as either benign or malignant. Feature reduction is also done along with feature extraction wherein the repetitions or unnecessary information in the data is being removed or filtered out. After feature extracting and feature reduction procedures, the next stage involves the classifications. Machine learning supports different classifiers such as Support Vector Machines, K-Nearest Neighbour based algorithms, decision trees etc. Based on the application of the corresponding classifier, different performance parameters could be predicted. Some of the parameters which are to be forecasted include accuracy, sensitivity and specificity. The abstract model for the implementations is shown in Fig.2.

The entire methodology is expanded on three levels. The initial level discusses the conceptual model consisting of the three levels of pre-processing, feature extraction and classification. The next level details the conceptual model along with the role of machine learning at every level. The pre processing involves the usage of filtering and data enhancements techniques. The extraction of features can be done using different techniques such as Principle Component Analysis, Genetic Algorithms or Wavelet Transforms. Finally different machine learning algorithms like Support vector Machines, K-Nearest Neighbour or decision trees could be used towards implementation of feature extraction. The third level of the abstract model involves the software implementation through different techniques. The common software platform generally to be used for the implementations is Python Language because of the ease of usage. The training and testing database initially proposed is to use the Wisconsin breast cancer database. The wide usage of this database is attributed to its characteristic features which include the large number of samples available and its unique feature of being noise-free. Further, there are only few missing values in this database. The proposed model is initially to be trained and tested using this existing database. Subsequently, it is proposed that a database is to be generated satisfying the pre-requisites of improved size of the dataset as well as increased accuracy in terms of providing noise free environment.

IV. RESULTS AND DISCUSSIONS

The database taken towards analysis is Wisconsin breast cancer database. The basic proposed model has been implemented using Bayes Classifier. Python programming language has been used towards developing software. The database has been developed by University of Wisconsin and this database has been utilized for performing the experimentation. There are different characteristic features which are present in the database. Some of them are thickness of the clump, uniformity with which cell sizes are present, different types of nuclei present etc. The general values associated with these characteristics change from 1 to 10. There are two different classes, to which these instances fall into which is either being benign or malignant.

Naive Bayes classifiers are probabilistic classifiers. These classifiers are based on Bayes' theorem. The accuracy of the proposed system is being calculated.
The basic formula towards finding accuracy is as follows:
Accuracy = (True Positive + True Negative)/ (True Positive + True Negative + False Positive+ False negative)

Table 1: Attributes and Accuracy

Number of attributes	Feature Sets	Accuracy
8	Thickness of lump, Uniformity of Cell Size, Uniformity of Cell Shape, Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses	97.96%

Totally for experimentation 8 attributes have been considered such as thickness of clump, cell size uniformity, cell size, type of nucleoli etc. Based on the experimentation done the accuracy is found to be 97.96%.
This outperforms the existing methods which are present in the literature.

V. CONCLUSION

Breast cancer results in lots of causality every year and hence there is worldwide research going on to mitigate the problem. There are many techniques being explored to process the voluminous amount of medical data present and detection of presence of an anomaly is a herculean task. There is huge requirement to process the data and create relevant datasets. Further, there is also growing needs to develop benchmark datasets which could provide the platform to create forecasting models. This paper comprehensively looks into various contributions in terms of building such predictive models for premature detection of breast cancer. Further, the paper proposes a wholesome framework built using machine learning techniques towards foretelling the presence of breast cancer. The usage of the machine learning algorithms along with the procedures for software implementation of the algorithms is detailed in the paper. Such structure when developed will greatly contribute to alleviate the increasing problems associated with detection and treatment of breast cancer.

REFERENCES

1. Ashwin Belle, Raghuram Thiagarajan,S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian, Big Data Analytics in Healthcare, BioMed Research International, Volume 2015, Article ID 370194, 16 pages, 2015.
2. Seth Earley, The Promise of Healthcare Analytics, Published by the IEEE Computer Society, 2015 IEEE, 1520-9202/15
3. A.Rishika Reddy, P. Suresh Kumar, Predictive Big Data Analytics in Healthcare, 2016 Second International Conference on Computational Intelligence & Communication Technology, 978-1-5090-0210-8/16 , 2016 IEEE , DOI 10.1109/CICT.2016.129
4. Van-Dai Ta, Chuan-Ming Liu, Goodwill Wandile Nkabinde, Big Data Stream Computing in Healthcare Real-Time Analytics, 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, 978-1-5090-2594-7116, 2016 IEEE
5. R. Vanathi, Dr. A. Shaik Abdul Khadir, A Robust Architectural Framework for Big Data Stream Computing in Personal Healthcare Real Time Analytics, World Congress on Computing and Communication Technologies (WCCCT), 978-1-5090-5573-9/16 2016 IEEE, DOI 10.1109/WCCCT.2016.32
6. Parag Chatterjee, Leandro J. Cymberknop, Ricardo L. Armentano, IoT-based Decision Support System for Intelligent Healthcare – Applied to Cardiovascular Diseases, 2017 7th International Conference on Communication Systems and Network Technologies, 978-1-5386-1860-8/17 , 2017 IEEE 362, DOI 10.1109/CSNT.2017.69
7. Ravi Kishore Kodali, Govinda Swamy and Boppana Lakshmi, An Implementation of IoT for Healthcare, 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | Trivandrum
8. Rajesh Jangade, Ritu Chauhan, Big Data with integrated cloud computing for Helathcare analytics, 978-9-3805-4421-2/16 ,2016 IEEE
9. Nina S. Godbole, John Lamb, Using Data Science & Big Data Analytics to Make Healthcare Green, 978-1-4673-7865-9/15, 2015 IEEE
10. Pawan Chowdhary, Sunhwan Lee, John Timm, Heiko Ludwig, Sarah Knoop, Coordinating Analytics Methods for Mobile Healthcare Applications, 2016 International Workshop on Software Engineering in Healthcare Systems, SEHS'16, May 14-15, 2016, Austin, TX, USA, 2016 ACM. ISBN 978-1-4503-4168-4/16/05. DOI: http://dx.doi.org/10.1145/2897683.2926711
11. Sunil Kumar and Maninder Singh, Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools, Big Data Mining And Analytics, ISSN 222096-0654 l l05/06l l, pp 48–57 Volume 2, Number 1, March 2019 DOI: 10.26599/BDMA.2018.9020031
12. C Imthyaz Sheriff, Twaseef Naqishbandi, Angelina Geetha, Healthcare Informatics and Analytics Framework, 2015 International Conference on Computer Communication and Informatics , *ICCCI* -2015
13. Cheryl Ann Alexander and Lidong Wang, Big Data Analytics in Heart Attack Prediction, 2017, 6:2, DOI: 10.4172/2167-1168.1000393
14. S.Suguna, Sakthi Sakunthala,N, S.Sanjana, S.S.Sanjhana, A Survey On Prediction Of Heart Diseases Using Big Data Algorithms, International Journal of Advanced Research in Computer Engineering & Technology , Volume 6, Issue 3, March 2017, ISSN: 2278 – 1323
15. Heba F. Rammal, Ahmed Z. Emam, Toward Robust Heart Failure Prediction Models Using Big Data Techniques, eTELEMED 2018 : The Tenth International Conference on eHealth, Telemedicine, and Social Medicine, IARIA, 2018. ISBN: 978-1-61208-618-7
16. Sang Hun Han , Kyoung Ok Kim, Eun Jong Cha, Kyung Ah Kim and Ho Sun Shon System Framework for Cardiovascular Disease Prediction Based on Big Data Technology, Symmetry 2017, 9, 293
17. Shaila H Koppad, Dr.Anupamma Kumar, Application of Big Data Analytics in Healthcare System to Predict COPD, 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], 2016 IEEE
18. Noppon Choosri, Krit Khwanngernb,Hongnian Yu, Krid Thongbunjobd Rattasit Sukhahutab,, Juggapong Natwichai , Pruet Boonma, ICT framework for collaborative healthcare services: A case study of Cleft Lip/Palate treatment network in northern Thailand, 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)
19. Santoshi Kumari, Haripriya.A, Aruna.A, Vidya.D.S, Nithya.M.N, Immunize - Baby Steps for smart healthcare Smart solutions to Child Vaccination, IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT'17), 2017 IEEE
20. Sangram Keshari Swain, Use Of Big Data Analytics In Lung Cancer Data Set, International Journal of Computational Engineering Research ,ISSN 2250 – 3005 Volume 07, Issue, 12 , December – 2017.
21. Amy Makler, Ramswamy Narayanan, Big Data Analytics and Cancer, MOJ Proteomics Bioinform 2016, 4(2): 00115
22. Big Data Analysis of Clinical records for Cancer Care,White Paper by Intel.
23. Ritu Parna Panda, Prakalpa Prakash Barik and 3P. Alok Kumar Prusty, A Review Paper on Big Data in Lung Cancer Big Data Analytics in Lung Cancer, International Journal of Trend in Research and Development, Volume 3(5), ISSN: 2394-9333
24. Desislava Ivanova, Big Data Analytics for Early Detection of Breast Cancer Based on Machine Learning, AIP Conference Proceedings 1910, 060016 , 2017
25. Savita Kumari Sheoran, Breast Cancer Classification using Big Data Approach, Paripex Indian Journal Of Research ,Volume 7, Issue 1, January 2018.
26. N. Thiebaut, A. Simoulin, K. Neuberger, I. Ibnouhsein, N. Bousquet, N.Reix, S. Molière and C. Mathelin, An innovative solution for breast cancer textual big data analysis

27. Umesh D. R. , B. Ramachandra, Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach, International Journal of Computer Applications (0975 – 8887) Volume 150, No.7, September 2016
28. L. Sankari,R. Rajbharath, Predicting Breast Cancer using Novel Approach in Data Analytics, International Journal of Engineering Research & Technology, ISSN: 2278-0181 ,Vol. 6 Issue 05, 2017
29. Gomathi, and Sandhya, Prognosis and Diagnosis of Breast Cancer Using Interactive Dashboard through Big Data Analytics, Biotechnology: An Indian Journal, Vol.13, Issue 1.
30. K. Shailaja, B. Seetharamulu, M.A. Jabbar , Prediction of Breast Cancer Using Big Data Analytics, International Journal of Engineering & Technology, 7 (4.6) (2018) 223-226
31. Minta Thomas, Kris De Brabanter, Johan AK Suykens and Bart De Moor , Predicting breast cancer using an expression values weighted clinical classifier, BMC Bioinformatics. 2014; 15(1): 411.
32. Vikas Chaurasia, Saurabh Pal and BB Tiwari, Prediction of benign and malignant breast cancer using data mining techniques, Journal of Algorithms & Computational Technology, 2018, Vol. 12(2) 119–126
33. K.Sivakami, Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model, International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-5, August 2015 ISSN: 2395-3470
34. Desta Mulatu, Rupali R. Gangarde, Survey of Data Mining Techniques for Prediction of Breast Cancer Recurrence, International Journal of Computer Science and Information Technologies, Vol. 8 (6) , 2017, 599-601
35. Rati Shukla, Vikash Yadav, Parashu Ram Pal, Pankaj Pathak, Machine Learning Techniques for Detecting and Predicting Breast Cancer, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7 May, 2019

AUTHORS PROFILE

Mithra Venkatesan completed her doctorate under Savitribai Phule Pune University, Pune, India in 2017. Currently, she is working in D.Y.Patil Institute of Technology, Pune as Associate Professor. With an overall teaching experience of over 15 years. She has worked as visiting scholar with Georgia Institute of Technology, Atlanta for 1 year. She has guided over 50 UG Projects. She has held various important portfolios during her tenure Her area of research includes Cognitive Radio, Artificial Intelligence and soft computing. She has published 30 papers in international conferences and reputed journals.

Anju VijayKumar Kulkarni completed her doctorate under University of Pune in 2008. Currently working as Professor of E&TC and Dean—R&D and Ph.D. in D. Y. Patil Institute of Technology, Pune, India with an overall teaching experience of over 30 years at Undergraduate and postgraduate level. She has guided more than 26 students towards post graduation and has guided seven research scholars towards Ph.D. She has more than 65 papers published in international conferences and reputed journal. She has held various important portfolios during her tenure. Her area of interest includes Wireless Networks Cognitive Radios, 5G technologies, Machine Learning and Pervasive Computing.

Radhika Menon is working as Professor in the Department of Mathematics. She is currently designated as Associate Dean - R & D at Dr. D. Y. Patil Institute of Engineering and Technology with an experience of around 20 years. Her areas of interest are Partial Differential Equations, CFD and Outcome Based Education. She is a registered guide at University of Pune (SPPU).

V.Ramakrishnan completed his postgraduate M.S from Indian Institute of Sciences, Bangalore. He has over 25 years experience in software services and product industry. His area of interest includes Artificial Intelligence, Machine Learning and Product Lifecycle Management.