Induction of Decision Trees based on Gray Wolf Optimizer for heart disease classification

Pravin S. Game, Vinod Vaze, Emmanuel M.

Abstract--- With the tremendous growth of the data and its usability to help in decision making, it has earned the status of asset today. This applies to medical data too. The decision trees is one of the most sought after tool in data mining for analyzing and representing the data for better visualization and to improve decision making. The deaths due to heart disease are at rise and hence the disease needs special attention not only in identifying that the person is suffering through a heart disease but also to identify which class of heart disease. This paper presents the use of gray wolf optimization to construct decision trees for classification of heart-disease data. The approach explored the opportunity to optimize the decision tree using the nature inspired gray wolf optimizer algorithm. Here in this work, decision trees are used to predict the heart disease. Standard heart disease dataset is used to validate the results. The experimental results are compared with the results from standard decision trees algorithms available in the data mining tool.

Keywords: Decision trees, Gray wolf optimizer, Heart disease, Induction of trees, Optimization.

1. INTRODUCTION

Heart disease is causing very large number of deaths across the globe. It is a leading cause of deaths in India, United States as well as in the world when compared with other diseases. This is corroborated by the multiple causes of deaths data available over years starting from 1977 till 2016 from Center for Disease Control and Prevention. One in four deaths is caused by heart disease. The reasons for the heart disease include the various medical conditions and lifestyle habits viz. excessive smoking and drinking, diabetic conditions, poor physical activity, improper diet, high levels of stress, high low-density lipoprotein cholesterol.

Let's look at the yearly mortality data [1] for US from 1999 till 2016, according to the classification provided by the International Classification of Diseases (ICD) in its tenth revision, presented in Table 1 and Fig 1. The ICD provides a list of causes of deaths. As per the 10th revision the diseases and corresponding chapters are elaborated after table 1. These are the major types or categories which are further sub-grouped into sub-types or categories specifying the reasons of death. If one observes the top 15 causes of deaths Table 2 and Fig. 2, it is clear that the heart disease related deaths in US tops the list. A similar report by WHO [2], where it identified 20 leading causes of death, heart disease tops the list.

The recent study by [3] observed that heart disease related deaths in India contribute to one-fifth of the deaths in the same category. This methodical study found that in 2015, 2.1 million deaths in India were due to cardiovascular diseases. A surprising finding was deaths due to ischemic heart disease- a type of cardiovascular disease- increased rapidly in rural areas and its rate surpassed the urban death rate due to same disease. The study concluded that to achieve the global goal of reducing non-communicable diseases by 2030, it is very important to make progress in controlling the deaths due to cardiovascular disease i.e. heart diseases.

Early detection of heart disease is of utmost importance. It is also pertinent to classify (identify) the type of heart disease, so that the proper medication can be prescribed. To classify a new patient showing symptoms of heart disease, the symptoms need to be converted to attributes. There is need to refer to a historical data having similar kind of attributes and a systematic classification done where the last attribute represents the category of the heart disease. Given the various ranges for each of the attributes, it becomes difficult to come to a conclusion and classify the disease manually. This is where machine learning will come to the assistance of the surgeons to quickly classify and identify the category of heart ailment of this new patient.

Machine learning is a field of computer science where computing model is trained using historical data and then this trained model is used for the destined purpose. There are two major methods of learning i.e. supervised and unsupervised. In supervised a set of input is mapped to an output. Whereas in unsupervised learning there is no such mapping and only input set is available. Majority of the machine learning is supervised. There are various approaches used in machine leaning to train the system. Of these, classification is one aspect which is the subject matter of this work. The well-known methods to build the learning models for classification include support vector machines, decision trees, random forest, linear regression, Naïve-Bayes, ensemble classifiers, neural networks [4].

Table 1: Number of Deaths in Each Year (1999-2016)

Year	Number of Deaths in US	Total Population in US
1999	5732491	279040168

Revised Manuscript Received on March 02, 2019.

Pravin S. Game, Research Scholar, JJT University, Jhunjhunu, Rajasthan, India
Dr. Vinod Vaze, JJT University, Jhunjhunu, Rajasthan, India
Dr. Emmanuel M., Pune Institute of Computer Technology, Pune, India

2000	5948173	281421906
2001	5986103	284968955
2002	6079340	287625193
2003	6153293	290107933
2004	6089036	292805298
2005	6293199	295516599
2006	6281454	298379912
2007	6308826	301231207
2008	6432432	304093966
2009	6347832	306771529
2010	6496415	308745538
2011	6715004	311591917
2012	6816497	313914040
2013	6919035	316128839
2014	6993189	318857056
2015	7284257	321418820
2016	7444681	323127513
Total	116321257	5435746389

ICD 10th revision has classified the diseases into 22 major categories in the respective chapters from chapter I to chapter XXII. Every chapter is then subdivided as per the specific disease (for example Cholera) which is further categorized according to cause (e.g. classical cholera). All the diseases are identified using codes (e.g. Cholera –A00). Here brief information on each chapter is provided as well the codes used are reproduced. Chapter I contain various communicable diseases which are infectious and parasitic in nature and codes used are from A00-B99. It covers diseases like cholera, typhoid, tuberculosis, plague, leprosy, syphilis, rabies, skin diseases, HIV. Chapter II provides the classification of diseases caused due to unnatural growth of cells- called as neoplastic diseases- including benign and malignant and coded from C00 to D48. Chapter III classifies anaemia caused due to various reasons and diseases related to immune system (D50-D89). Chapter IV has diseases related to thyroid and endocrine glands, diabetes, obesity, and metabolism (E00-E90). Next chapter has classification of metal diseases (F00-F99). Chapter VI covers the disorders of nervous system (G00-G99) and the next covers disorders of eye (H00-H59) and the next identifies ear disease (H60-H93). Various heart diseases and circulatory system disorders are classified in chapter IX (I00-I99). Respiratory diseases and digestive disorders are categorized in chapters X and XI respectively. Subsequent chapters classify skin diseases (L00-L98), tissue, muscle and skeletal system diseases (M00-M99), genital and urinary diseases (N00-N98), pregnancy and childbirth related (O00-O99 and P00-P96), congenital diseases (Q00-Q99), and injury, poisoning and other external causes (S00-T98, V01-V98) accordingly. It has also classified diseases caused due to infliction of injury- because of accidents, self-harm and assault; and due to health services.

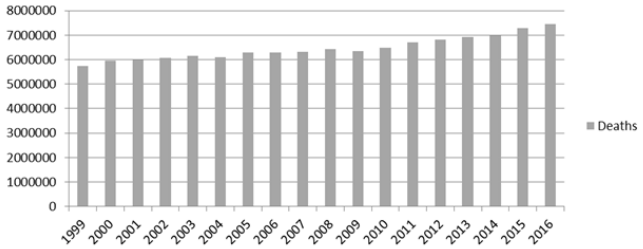

Fig. 2: Deaths due to 15 leading causes for the period starting from 1999 till 2016

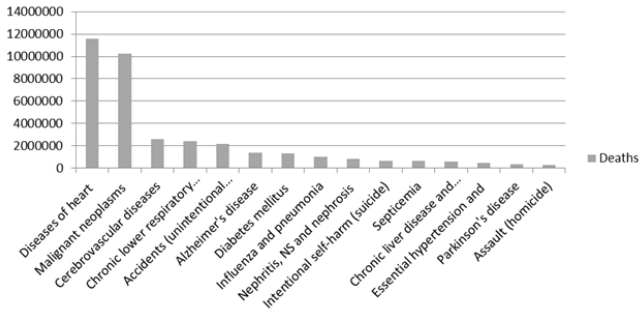

Table 2: 15 Leading Causes of Deaths (Period 1999-2016)

UCD - 15 Top Causes leading to Death	Deaths
Heart diseases (Rheumatic, Ischemic, Pulmonary, Nonrheumatic, Hypertensive, others)	11575183
Neoplasms (Malignant)	10244536
Cerebrovascular related – Haemorrhage, stroke, infraction	2580140
Bronchitis, Asthma, Emphysema	2434726
Accidents, Assault by drugs, assault by corrosive substance (unintentional injuries)	2177884
Alzheimer's disease	1373412
Diabetes: type 1, type 2, malnutrition related, specified and unspecified	1316379
Pneumonia, Pandemic and seasonal influenza	1038969
Nephritis, renal failure, contracted kidney	807980
Intentional self-harm (harm by poisoning, hanging or other measures)	649843
Septicemia- streptococcal, staphylococcal	635097
liver disease: cirrhosis, alcoholic, fibrosis	562382
Essential and secondary hypertension, and hypertensive renal disease	453805
Parkinson and secondary Parkinson	377956
Assault self-harm (homicide)	314705

Each of these methods has its own advantages and disadvantages. The choice of the algorithm is most of the times the interest of the researchers. Our interest is decision trees. Many researchers have also suggested use of hybrid approaches for classification. Use of one such hybrid approach is presented here. The paper is organized as

follows: relevant literature survey is presented in section 2, followed by hybrid approach in section 3. Results are discussed in section 4 followed by conclusion in section 5.

2. LITERATURE SURVEY

The computer science has been at the forefront in using the algorithms which mimic the behaviors of other living entities. The artificial intelligence is full of such algorithms and is being continuously updated to date. Here three such nature inspired algorithms are briefed first and then their recent variations and applications including for disease classification are discussed.

Recently, based on the natural living of the grey wolf, an optimization algorithm was proposed by [5] called as grey wolf optimizer (GWO). This algorithm has been extensively being used in various fields to optimize the results. The algorithm considers the strict hierarchy in the social order and the way these wolves hunt the pray. Socially the grey wolves are of four categories viz. alpha (α), beta(β), delta(δ) and omega(ω). Alpha are the leaders and the decision makers in the pack of 5-12. Alpha is a male and a female. Their decision is to be followed by the remaining. Next are the beta wolves, a male and a female, which are subordinate to alpha and make sure that the decision of alpha is followed by the others. Omega are the lowest in the hierarchy and are the last one to eat. All remaining are delta wolves following the orders of alpha and beta. These four categories can be used to represent the quality of the results. When hunting their movement can be mapped to the direction in which the solution moves from feasible to optimal solution.

Other widely used algorithms are genetic algorithms (GA) and particle swarm optimization (PSO). Genetic algorithms are based on the evolutionary principle of 'survival of the fittest'. In nature the evolution happens with the change in the genetic patterns in such a way that it helps the organism to survive in the nature in a better way. This is generally called as adaptive nature or evolution. This is mathematically modelled by [6] so that it can be used in artificial intelligence and other complex systems. An initial population is randomly selected. Algorithm uses three operators for modifying the gene- selection, crossover and mutation. The new gene selected is the one maximizing the profit or leading to the optimal solution. The algorithm is iterative and keeps on iterating till no further maximization is possible. To analogize with natural selection, the algorithm stops when no better gene could be created. Subsequently in [7], the use of genetic algorithms based approaches for classification in machine learning is briefed. In the handbook [8], the applications of genetic algorithm to solve various complex problems are described.

Particle swarm optimization [9] is based on the social behaviour of birds or fishes. The bird flocks or the fish schools have a natural way of functioning making it sure that that they don't collide with each other by adjusting the velocity and direction, that they avoid their predators, that they get food and mates. It also proceeds in similar manner to GA by initializing a sample random population representative of solutions. Each solution is considered as a non-colliding fish –termed as particle- and let it swim through the hyperspace of solutions to improve fitness. Unlike GA, PSO does not any operators making it simple for implementation. The particle keeps storing the fitness values achieved till now- called as pbest, for global best solution obtained by any particle is called as gbest and the best solution achieved in local environment is called lbest.

Cancer is caused by some genes and hence all genes don't contribute to its occurrence. So the issue is, from the given large number of genes identify such informative genes. For such analysis gene expression datasets are available which needs to be classified in the category of cancer causing genes and not causing genes. PSO and GA have been extensively used for this purpose. One such approach is presented in [10], wherein PSO is used along with well-known C4.5 decision tree for classification. The approach is called as PSODT. PSO initializes a population function as calculates the fitness of each particle using the accuracy of C4.5 algorithm. The local as well as global solutions are updated till the global solution value is not met. The method is applied to 11 standard datasets to validate and is compared with standard methods. The PSODT was found to be better than all the compared methods for cancer classification.

A hybrid approach using PSO and GA was proposed in [11] in which support vector machine (SVM) was used for classification of gene expression data. Here the operators of GA are integrated in PSO. First population is generated, fitness is calculated and then PSO operators are applied to find final solution. However, if PSO terminates with the solution not satisfying the termination condition, GA is used to find the final solution. If GA also terminates without giving the solution satisfying the termination condition, the steps from application of PSO operators are repeated. The iterations are repeated for predefined number of times. SVM is used for classification. The implementation results found to improve the accuracy of classification with this hybrid PSO/GA approach. A variant of PSO called enhanced binary particle swarm optimization (EPSO) was proposed in [12] with the purpose to select most relevant subset of informative genes. EPSO differed from earlier variants of PSO in the sense; it added particle speed as a characteristic of particle. On experimenting with 10 cancer datasets, the EPSO could find smaller subsets of genes, helping in better classification results. Another approach for cancer classification, in which combination of PSO with SVM and combination of GA with SVM are deliberated [13], [14].

For heart disease prediction, a model using PSO with K-nearest neighbour (KNN) classifier is proposed in [15], wherein PSO is used to for feature selection to improve the accuracy. The combination of PSO with KNN gave better results. However, when this combination is integrated with interquartile range, the model gave 100% accuracy for the classification of heart-disease dataset with 12 attributes and 97.5% accuracy for dataset with 10 attributes.

GA based optimization approach was proposed by [16], where GA was used for feature selection. The authors studied the standard classification approaches viz. random forest, naïve Bayes, decision tree, and SVM. After completing the pre-processing of data, GA was applied to

select best features. Then k-fold cross-validation is applied to generate train and test data. Training data is then used for learning by the respective algorithms and the training data is used for validation. This approach found improvements in accuracy, sensitivity, specificity and area under curve (AUC) measures of the classifiers used for classifying four heart-disease datasets.

Classification system for heart disease using PSO and naïve Bayes was proposed in [17], wherein, the PSO was used to select the best features. The features with low value of PSO were rejected. Then the naïve Bayes was applied to the selected best features. The experiment is done with changing the maximum number of iterations from 10 to 100 and the selected features were recorded. These selected features varied in different iterations to some extent. After applying the NB the classification accuracy recorded showed significant improvement than the compared systems. The classification accuracy of NB-PSO was recorded as 87.91%, which was 8.79% more than the accuracy of the compared systems. However, the combination of NB with GA for the same dataset showed 86.29% accuracy, which was comparable with the NB-PSO approach.

Another classification system using NB and GWO was proposed in [18], called as GWO-NB for heart disease classification. Before proceeding for classification, the data is discretized using class-attribute interdependence maximization approach. This discretized data is then divided into training and testing data. Training data is optimized using GWO to identify the attributes with best weights. These attributes are then used to construct the classification model based on NB. Once this GWO-NB model is ready, it is then fed with the testing data to validate the functioning, using the chosen best attributes for classification. The system is evaluated using standard heart disease dataset and accuracy is found to be 87.45% with discretization and 85.79% without discretization. The accuracy results are compared with 12 existing methods and the results showed the improvement ranging from 2 to 28% compared with these existing system.

3. PROPOSED SYSTEM

In this work grey wolf optimizer [5] is used to identify the best features from the heart disease dataset. These are then provided to the standard Iterative Dichotomiser-3(ID3) [19] for classification. The dataset considered is VA Long Beach heart disease dataset from UCI [20]. The dataset has 14 attributes (last attribute is the class) and 200 instances. The instances are classified into 5 classes.

The steps are as follows:

1. The Long Beach VA dataset is provided as input.
2. GWO is applied to select best attributes.
3. Based on these attributes, model for ID3 is constructed.

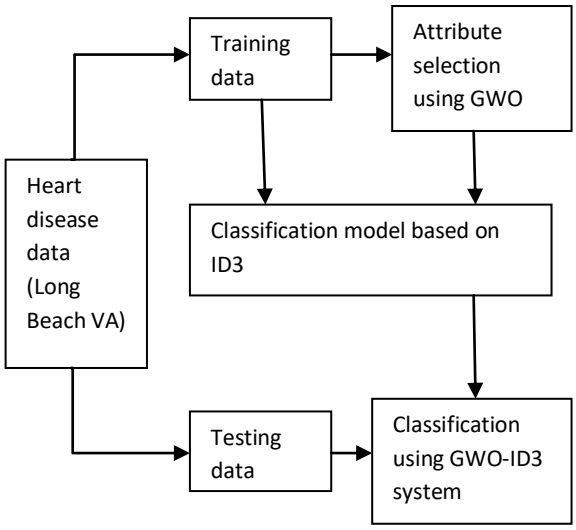

Fig. 3: Classification system using GWO-ID3

4. Now, the GWO-ID3 system applied to test data.
5. Evaluate the performance of the system.

The schematic representation is given in Fig. 3.

4. RESULTS

The system is tested for accuracy of the classification by measuring correctly classified and incorrectly classified instances. The proposed system is implemented in Java and the standard methods available in Weka tool for classification are compared for accuracy with the proposed system. Table 3 presents the results for Long Beach VA heart disease dataset with 10-fold cross-validation. Table 4 shows the results for 70% training data and 30% testing data.

Table 3: Accuracy with 10-fold cross-validation

Algorithm	Correctly classified instances	Incorrectly classified instances
Hoeffding tree	53	147
J48	58	142
Logistic Model Trees	66	134
Naïve Bayes	57	143
Random Forest	71	129
Random Tree	65	135
SimpleCART	59	141
GWO-ID3	55	145

Table 4: Accuracy with 70-30 split of data

Algorithm	Correctly classified instances	Incorrectly classified instances
Hoeffding tree	11	49
J48	15	45
Logistic Model Trees	10	50
Naïve Bayes	14	46
Random Forest	18	42

Random Tree	13	47
SimpleCART	15	45
GWO-ID3	14	46

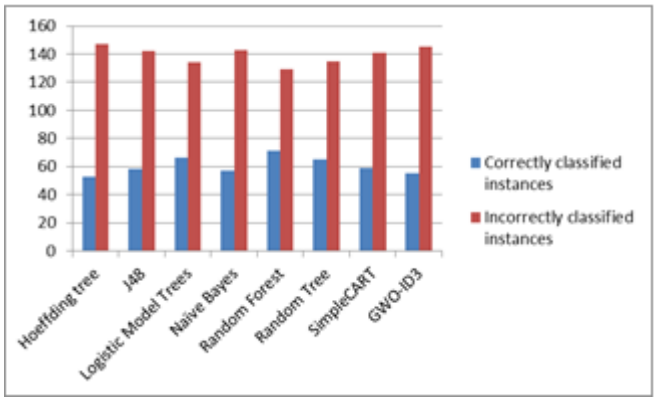

Fig. 4: Comparison of accuracy when 10-fold cross-validation is used

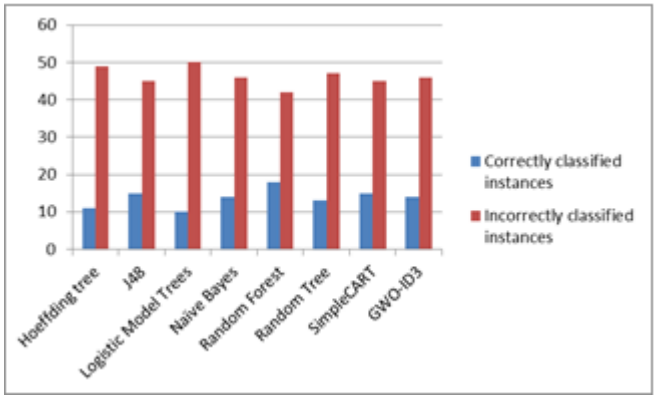

Fig. 5: Comparison of accuracy when 70-30 spitting is used

5. CONCLUSION

Heart disease classification is a crucial problem as this disease is the causing maximum deaths followed by cancer. The paper has discussed the gravity of situation by presenting the disease related data. The proposed method in which GWO is combined with ID3 to classify the Long Beach VA dataset is found to give comparable results as can be seen from the Fig. 4 and Fig. 5. It gave better results than hoeffding trees, logistic model trees, random trees when splitting of dataset is used. Whereas in case of 10-fold cross validation, it classified better than Hoeffding tree only. This also suggests that the model can be enhanced to improve the performance. In future, some preprocessing methods will be used for better selection of attributes to improve the classification accuracy.

6. REFERENCES

1. CDCP, National Center for Health Statistics. Multiple Causes of Death database, for 1999-2016, released in December, 2017. Data collected and compiled through Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10.html on Nov 16, 2018
2. World Health Organization, Global Health Estimates: 20 leading causes of death, 2000 and 2015, December 2016, http://www.who.int/healthinfo/global_burden_disease/en/ Accessed on Nov 16, 2018.
3. Ke C, Gupta R, Xavier D, Prabhakaran D, Mathur P, Kalkonde YV, Kolpak P, Suraweera W, & Jha P (2018), "Divergent trends in ischaemic heart disease and stroke mortality in India from 2000 to 2015: a nationally representative mortality study", *Lancet Global Health* 6, e914-923.
4. Cady F (2018), The Data Science Handbook, 1st edn., *Wiley India Pvt. Ltd*., pp. 99-113.
5. Mirjalili S, Mirjalili AM, & Lewis A (2014), Grey wolf optimizer, *Adv. in Engg. Software* 69, 46-61.
6. Holland JH, Adaptations in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence, MIT Press, (1992).
7. Goldberg DE, Holland JH, (1988) "Genetic Algorithms and machine learning", *Machine Learning*, Vol. 3, issue 2-3, (1988), pp. 95-99.
8. Chambers L, Ed., *The practical handbook of Genetic Algorithms*, (200), Chapman & Hall, CRC, 2nd Edn. Available online.
9. Kennedy J, Eberhart R, (1995), "A new optimizer using particle swarm theory", *6th Inter.Symp. on Micro Machine and Human Sci.*, Nagoya, Japan, pp. 39-45.
10. Li S, Wu X, & Tan M, (2008), "Gene selection using hybrid particle swarm optimization and genetic algorithm", *Soft Comp.*12, issue 11, pp. 1039-1048.
11. Chen KH, Wang KJ, Tsai ML, Wang KM, Adrian AM, Cheng WC, Yang T, Teng N, Tan K, & Chang K, (2014) "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm", *BMC Bioinfo.*, vol. 15, no. 49.
12. Mohamad MS, Omatu S, Deris S, Yoshioka M, Abdullah A, & Ibrahim Z,(2013), "An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes", *Algo. for Molecular Bio.*, vol. 8, no. 15. Available online last accessed 18 November 2018.
13. Huerta E, Duval B, & Hao JK, (2006), "A hybrid GA/SVM approach for gene selection and classification of microarray data," *Appl. of Evolutionary Comp.*, pp. 34-44.
14. Alba E, Garcia-Nieto J, Jourdan L, & Talbi EG (2007), "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms," *IEEE Congress on Evolutionary Computation* (CEC 2007), pp. 284-290.
15. Jabbar MA, (2017), "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", *Biomedical Research*, vol. 28, no. 9, pp. 4154-4158.
16. Suresh P, Anandraj MD, (2018), "Study and Analysis of Prediction Model for Heart Disease: An Optimization Approach using Genetic Algorithm", *Inter. J. of Pure and Applied Mathematics*, vol. 119, no. 16, pp. 5323-5326.
17. Dulhare UM, (2018), "Prediction system for heart disease using Naive Bayes and particle swarm optimization", *Biomedical Research*, vol. 29, no. 12, pp. 2646-2649.
18. El Bakrawy LM, (2017), "Grey Wolf Optimization and Naive Bayes classifier Incorporation for Heart Disease Diagnosis", *Australian Journal of Basic and Applied Sciences*, vol. 11, no. 7, pp. 64-70.
19. Quinlan JR, (1986), "Induction of decision trees", *Machine Learning*, vol. 1, pp. 81-106.
20. Janosi A, Steinbrunn W, Pfiesterer M, & Detrano R, Heart Disease data set. UCI Machine learning repository. Accessed 20 September 2018.

Pravin S. Game is a Ph.D. scholar in Computer Engineering at Shri JJT University, Jhunjhunu, Rajasthan. He received his Master of Engineering in Computer Engineering from Savitribai Phule Pune University and Bachelor of Engineering in Computer Science & Engineering from Sant Gadge Baba Amravati University. Currently, he is working at Pune Institute of Computer Technology, Pune. His research interests include data mining, big data analysis, machine learning.
Email:pravinsgame@gmail.com, psgame@pict.edu

Dr. Vinod Vaze is Ph.D. in Computer Engineering. He is B. Tech. from I.I.T., Kanpur, and has also earned PGDFM, Diploma in Cyber Law. He is currently working in Department of Computer Science and Engineering at Shri JJT University, Jhunjhunu, Rajashtan. His research interest includes machine learning, and cyber security.

Dr. Emmanuel M. is Ph.D. in Computer Science and Engineering. He is M. Tech. and B. Tech. in Computer Science and Engineering. He is currently working at Pune Institute of Computer Technology, Pune. He is leading Big Data Research group at PICT. His research interest includes big data, business intelligence, medical image processing and machine learning.