IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Soil Fertility Prediction for Yield Productivity and Identifying the Hidden Factors through Machine Learning Algorithms

R. Jayalakshmi[1*], M. Savitha Devi[2]

[1]Department of Computer Science, Sri Vidya Mandir Arts and Science College, Uthangarai, Krishnagiri, India
[2] Department of Computer Science, Periyar University Constituent College of Arts and Science, Harur, Dharmapuri, India

**Corresponding author: jayasvm179@gmail.com. Tel: +91 9791650170*

Available online at: www.ijcseonline.org

Accepted: 16/Jan/2019, Published: 31/Jan/2019

Abstract— Data mining is a promising technology which helps to analyze the data and to discover the interesting hidden patterns in large volume of data. The goal of data mining is to predict, identify, classify and optimize the use of resources to recognize complex patterns and make intelligent decisions based on data. Agriculture plays a vital role in economy and it is the backbone of our economic system. Data mining in agriculture provides many opportunities for exploring hidden patterns in these collections of data. Soil Fertility is the capability of soil to provide plants with enough nutrients and moisture to yield crop in better way. The yielding capability of a soil depends on soil fertility. It is very important to achieve and maintain an appropriate level of soil fertility for crop production. The main focus of this paper is to analyse the soil data which is collected from soil testing laboratory and identifying attributes to predict fertility from collected dataset by using different Machine Learning algorithms. This work also focuses on finding the best classification algorithm based on accuracy and performance measure using the soil dataset with different Data Mining classifiers like J48, Naïve Bayes and REPTree.

Keywords— Agriculture, Classification, Data Mining, J48, Naïve Bayes, REPTree, Soil fertility

I. INTRODUCTION

Data mining involves the extraction of implicit, interesting information from a database. Classification is an important Data mining's machine learning technique which is used to predict data instances from dataset. It involves the order wise analysis of large amount of informations. Data mining applications are used in various areas such as health care, insurance, medicines, Agriculture, banking and soil management. Agriculture is a business with risk that depends on climate, geography, political and economic factors. The integration of Computer Science with agriculture will generate new emission in management of agriculture information and it helps the farmers to make critical farming decision. Data Mining in Agriculture is a very recent research topic. Various data mining techniques are used in the field of agriculture that helps in prediction of problems, disease detection, optimizing the pesticide and so on. Nowadays, huge information on agricultural related activities are provided by recent technologies and this can be analyzed in order to find important information. Some of the factors that influence soil fertility are,

- Soil structure
- Minerals present in the soil
- Acidity or soil PH
- Content of organic matter
- Water draining ability of the soil
- Nutrient release capability
- Present soil characteristics
- Ground water
- Active soil life
- Infiltration of water

Soil fertility prediction plays a vital role in agriculture particularly food production. The main objective of this work is to identify an attributes for soil fertility and predict particular crop using the soil dataset.

The paper is organized as follows, Section II contains the introduction of Soil Fertility, Section III contains the Review of Literature, Section IV explains the Proposed Methodology with flow chart, section V explains about data collection, Section VI describes Experimental Analysis, Section VII discusses the various Data mining classification models, Section VIII describes results and discussion, and Section IX concludes research work with future directions.

II. SOIL FERTILITY

Soil fertility is an ability of the soil to supply essential plant nutrients and soil water in adequate amounts and proportions for plant growth and reproduction in the absence of toxic substances which may inhibit plant growth.

Soils are composed of five main components:

- Mineral particles derived from rocks by weathering
- Organic materials - humus from dead and decaying plant material
- Soil water - in which nutrient elements are dissolved
- Soil air - both carbon dioxide and oxygen
- Living organisms including bacteria that help plant decomposition.

The major and micro or trace elements are made available to plants by breakdown of the mineral and organic matter in the soil. Availability of these nutrients depends on how much is present, the form in which it is present in the soil, the rate at which it is released from organic matter or mineral particles and the soil pH i.e. its acidity or alkalinity.

The nutrients of soil are classified into three categories as shown in Table 1. Mineral ratios are also depending on type of soil available in that region.

Table 1. Plant Nutrients in Soil

Primary(Major)	Secondary	Trace(Minor)
Nitrogen(N) Phosphorus(P) Potassium(K)	Calcium(Ca) Magnesium(Mg) Sulphur(S)	Boron(B) Iron(Fe) Manganese(Mn) Copper(Cu) Zinc(Zn) Molybdenum(Mo) Chlorine(Cl)

III. REVIEW OF LITERTURE

Manisha Sahane, Balaji Aglave, Razaullah Khan, Sanjay Sirsat [1], describes the discussion on the application of data mining predictive and descriptive techniques in the field of agriculture. This information will help to select type of fertilizer, time of application and overall selection of cropping pattern and cropping system.

Dr.S.Hari Ganesh, Mrs. Jayasudha [2], They suggested an analysis of the soil information using completely different algorithms and prediction technique and also demonstrated a comparative study of varied classification algorithms i.e.Naïvebayes, J48 (C4.5), JRip with the assistance of data mining tool. J48 is incredibly easy classifier to form a decision tree.

Hetal Patel, DharmendraPatel [3], provide a brief review of a variety of Data Mining techniques that have been applied to model data from or about the agricultural domain. It also focuses on different data mining applications in solving the different agricultural problems, so it is useful for researchers to get information of current scenario of data mining techniques and applications in context to agriculture field.

P. Jasmine Sheela, K. Sivaranjani [4], uses different data mining techniques to analyze the soil test data sets that are collected from the authority. The former researchers used data mining techniques to build the automated system for soil datasets.

Vrushali Bhuyar [5] made an analysis of the soil fertility prediction using random forest classifier, naive bayes, J48 and compares the performance of these algorithms based on the fertility index. J48 algorithm used to calculate the soil fertility rate based on the nutrients found in the soil sample different fertilizers can be needed.

Jay Gholap, Anurag Ingole, JayeshGhoil, Shailesh Gargade, Vahida Attar [7] made a comparative study of classification algorithms such as JRip, Naïve Bayes, and J48. These classification algorithms implemented using WEKA tool. J48 gives the best accuracy level is 91.90%. This researcher builds an automated system for classifying soils easily.

IV. PROPOSED METHODOLOGY

The objective of proposed work is to analyze the agriculture data using data mining techniques. It will also help in determining what fertilizer to use for the crop and amount of fertilizer required. The methodology starts with soil data collection; later pre-processing the dataset has been carried out to filter missing attributes values, noisy data and miss match. They are filtered using WEKA filters then Data Mining classification followed by the evaluation of results. Framework of proposed methodology as shown in the Figure.1.

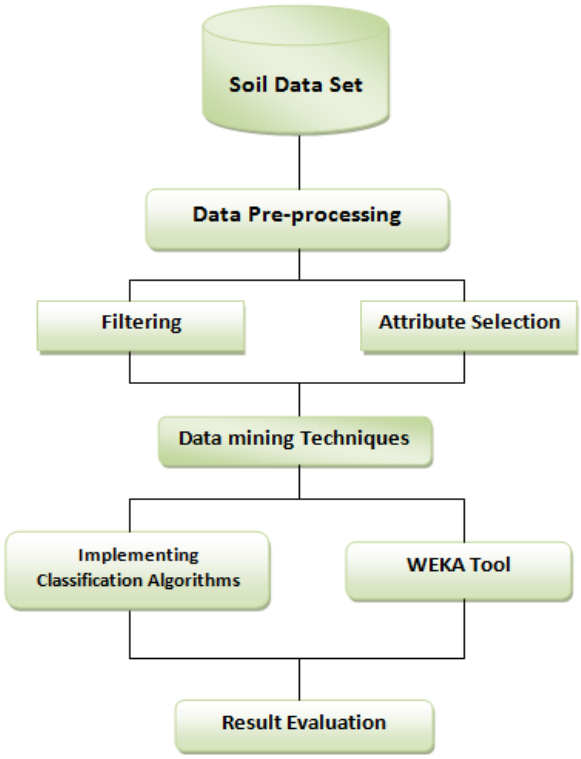


Figure.1 Framework of Proposed methodology

V. DATA COLLECTION

The dataset for this research is obtained on the basis of collecting soil samples from different irrigated regions of Vellore District, Tamilnadu through soil testing laboratory, Melalathur, Gudiyathum, to infer the range of soil parameter supporting healthy growth of plants.

VI. EXPERIMENTAL ANALYSIS

A. Data Mining with WEKA

The WEKA 3.9 (Waikato Environment for Knowledge Analysis) workbench is a state of art for machine learning algorithms and data pre-processing tools. It is the open source software for Data Mining. It is issued under the GNU, General Public License. It is fully implemented in Java programming language, so it is portable. Also it includes algorithms for Classification, Clustering, Association Rule Mining, Regression and Attribute Selection.

B. Attribute Selection

The Data miner tool supports many in-built Machine Learning algorithms. We have applied one of the filter methods under supervised option, because Classification comes under Supervised Learning Method. For this work the dataset with 100 records has been created in Excel 2007 and later it has been saved in the format of .CSV. Later the CSV formatted dataset has been opened and saved in the format of ARFF which is accepted file format for our mining process. The attributes taken for the dataset are listed below:-

Table.2 Attributes taken for work

S.No	Attributes	Description
1	Soil texture	Type of soil
2	PH	pH value of soil
3	EC	Electrical conductivity
4	OC	Organic Carbon
5	N	Nitrogen
6	P	Phosphorous
7	K	Potassium
8	Ca	Calcium
9	Mg	Magnesium
10	S	Sulphur
11	B	Boron
12	Cu	Copper
13	Fe	Iron
14	Mn	Manganese
15	Zn	Zinc
16	Na	Sodium

Among 16 attributes, 10 attributes were selected using select attribute option. For this attribute selection process, we have chosen InfoGainAttributeEval as Attribute Evaluator and Ranker as Search Method. The attributes selected for this work are PH, EC, OC, N, P, K, Fe, Zn, Mn, Cu.

VII. DATA MINING CLASSIFICATION MODELS

Soil classification was measured seriously to review, because of relying, upon the fertility class of the soil domain information, consultants determines that crops ought to be taken on it specific soil and that fertilizers ought to be used for an equivalent. The subsequent section describes Naive bayes, J48, JRip algorithms in short.

WEKA supports number of Classification algorithms. One of the main benefits of its platform is supporting Machine learning algorithms for our machine learning problems. The classification algorithms used for this work will be discussed here.

A. J48

J48 is an open source Java implementation of the C4.5 algorithm in the Data Mining tool, where J for Java and 48 for C4.8, hence J48 name. It is a minor extension of the C4.5 algorithm. C4.5 generates the decision tree used for classification. It builds a decision trees from a set of training data. At each node of decision tree, C4.5 chooses the attributes of the data that effectively splits its sets of sample into subset enriched in one class or the other. This splitting is the normalized information gain and the attribute with the highest normalized information gain is chosen to make the decisions.

In this proposed work, J48 Classifiers performs better compared to other classification algorithms such as Naïve Bayes and REPTree.

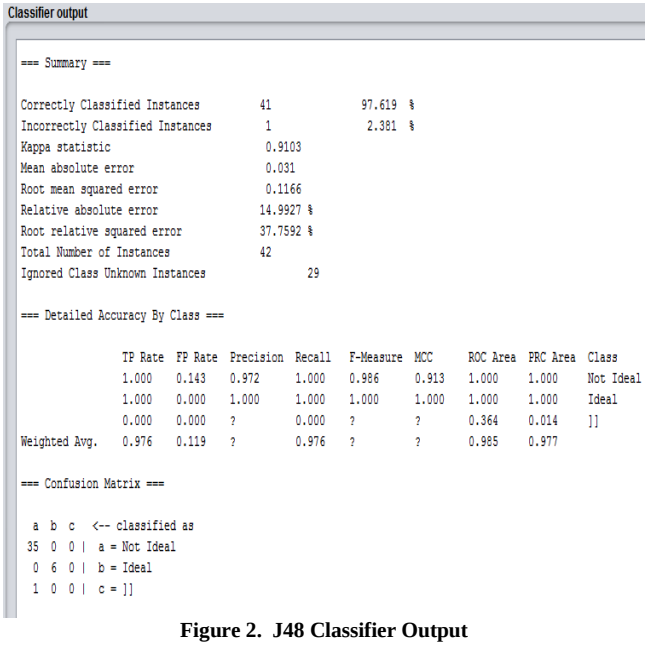


Figure 2. J48 Classifier Output

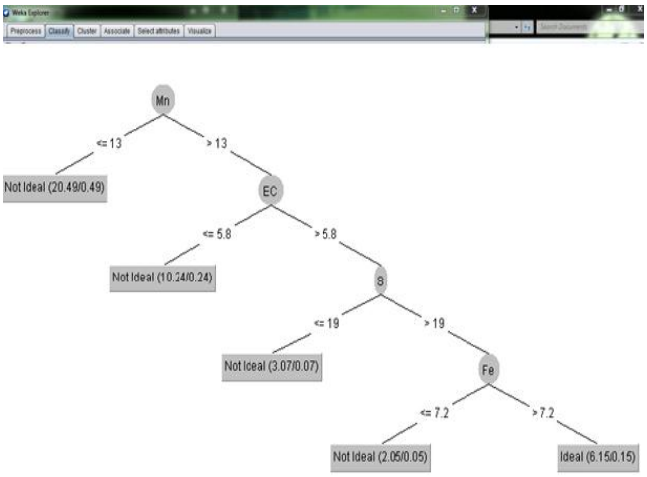
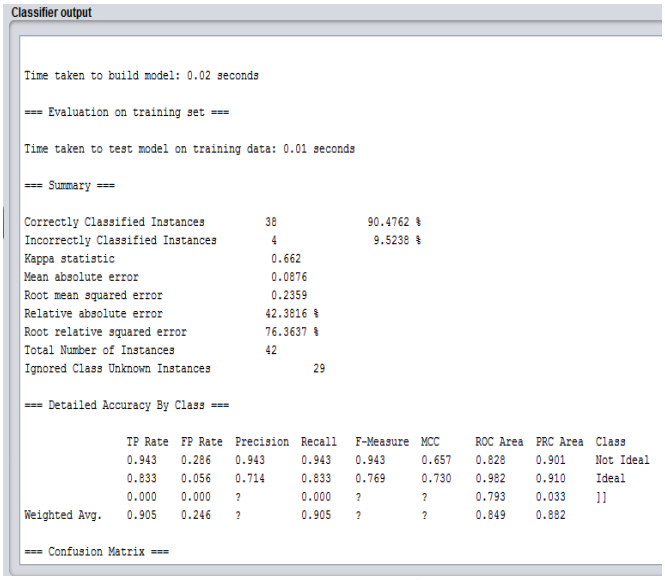


Figure 3. Visualize tree generated through J48 Classifier

B. Naïve Bayes

In Machine Learning, Naïve Bayes is not a single algorithm but a family of Classification algorithms based on Bayes rule of conditional probability. It analyses the data individually for their dependency as well as the independency among each other by making use of all the attributes in the dataset.



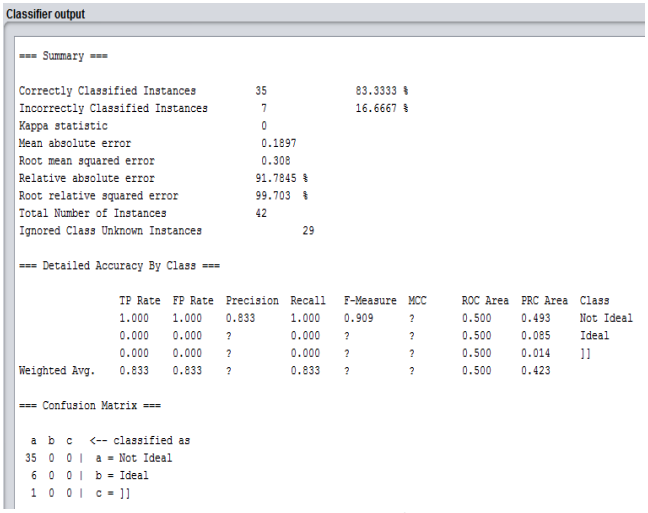
Classifier output

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          38               90.4762 %
Incorrectly Classified Instances         4                9.5238 %
Kappa statistic                          0.662
Mean absolute error                      0.0876
Root mean squared error                  0.2359
Relative absolute error                 42.3816 %
Root relative squared error             76.3637 %
Total Number of Instances               42
Ignored Class Unknown Instances                 29

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.943    0.286    0.943      0.943    0.943      0.657    0.828     0.901     Not Ideal
               0.833    0.056    0.714      0.833    0.769      0.730    0.982     0.910     Ideal
               0.000    0.000    ?          0.000    ?          ?        0.793     0.033     ]]
Weighted Avg.  0.905    0.246    ?          0.905    ?          ?        0.849     0.882

=== Confusion Matrix ===
```

Figure 4. Naïve Bayes Classifier Output

C. REPTree

REPTree algorithm is the fast decision tree learner which uses the regression tree logic and builds multiple trees. Later it selects best one among the generated trees. It is based on C4.5 algorithm which can produce classification or regression trees. It generates decision tree using information gain/variance. It prunes it using reduced error pruning.



Classifier output

```
=== Summary ===

Correctly Classified Instances          35               83.3333 %
Incorrectly Classified Instances         7               16.6667 %
Kappa statistic                          0
Mean absolute error                      0.1897
Root mean squared error                  0.308
Relative absolute error                 91.7845 %
Root relative squared error             99.703  %
Total Number of Instances               42
Ignored Class Unknown Instances                 29

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    1.000    0.833      1.000    0.909      ?        0.500     0.493     Not Ideal
               0.000    0.000    ?          0.000    ?          ?        0.500     0.085     Ideal
               0.000    0.000    ?          0.000    ?          ?        0.500     0.014     ]]
Weighted Avg.  0.833    0.833    ?          0.833    ?          ?        0.500     0.423

=== Confusion Matrix ===

  a  b  c   <-- classified as
 35  0  0 |  a = Not Ideal
  6  0  0 |  b = Ideal
  1  0  0 |  c = ]]
```

Figure 5. REPTree Classifier Output

Using three classification algorithms J48, Naïve Bayes and REPTree, the dataset taken for this research work has been tested and analyzed. Also comparisons among these algorithms have been done and concluded that J48 is best among the implemented algorithms.

VIII. RESULT AND DISCUSSION

In this paper, different classifier techniques of data mining are compared and evaluated on the basis of Mean absolute error rate, accuracy and time taken to build a model. Table 3. Shows mean absolute error and an accuracy rate of different classifier.

Table 3. Error and Accuracy Rate of different classifiers

Classifier	Mean absolute error	Accuracy rate
J48	0.031	97.6%
Naïve Bayes	0.0876	90.5%
REPTree	0.1897	83.3%

This experiment shows that J48 model is the best classifier technique, which achieves higher performance with low mean absolute error of 0.031 and accuracy 97.6 for soil fertility prediction.

IX. CONCLUSION AND FUTURE WORK

In this research paper, the soil data collected from soil testing laboratory, Melalathur, Gudiyathum, Vellore district has analyzed and predicted the range of main attributes that best suited for corn crop by using certain classifiers J48, Naïve Bayes and REPTree, Random Forest Also the comparison among these classifiers has been done and J48 classifier is proved to be the best among these algorithms. The future work of this research is to implement this in other Data Mining Tool R to predict the soil fertility for more crops by using extended dataset with high accuracy.

ACKNOWLEDGMENT

The Authors would like to thank, Senior Agricultural officer, Soil testing Laboratory, Melalathur, Gudiyathum for his cooperation and providing data for this research.

REFERENCES

[1] Manisha Sahane, BalajiAglave, Razaullah Khan, Sanjay Sirsat, *An Overview of DataMining Techniques Agricultural Soil Data Applied to Agricultural Soil Data*,International Journal of Agriculture Innovations and Research, Vol.3, No.2,pp. 445 – 448, 2014.

[2] Dr.S.Hari Ganesh, Mrs. Jayasudha, *an Enhanced Technique to Predict the Accuracy of Soil Fertility in Agricultural Mining*, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue. 7, pp. 285-287, 2015.

[3] Hetal Patel, Dharmendra Patel, *A Brief survey of Data Mining Techniques Applied to Agricultural Data*, International Journal of Computer Applications (0975 – 8887) Vol. 95, No. 9, pp. 6-8, 2014.

[4] P. Jasmine Sheela, K. Sivaranjani, *A Brief Survey of Classification Techniques Applied To Soil Fertility Prediction*, International Conference on Engineering Trends and Science &Humanities (ICETSH-2015), Vol. 3, No. 5, pp. 80-83, 2015.

[5] VrushaliBhuyar, *Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad district*, IJETTCS International Journal of Emerging Trends & Technology in Computer Science Issues, Vol. 3, Issues. 2, pp.200-203, 2014.

[6] Jay Gholap, *Performance Tuning of J48 Algorithm for Prediction of Soil Fertility*, Asian Journal of Computer Science and Information Technology, Vol.2, Issues 8, pp. 251-252, 2012.

[7] Jay Gholap, Anurag Ingole, JayeshGhoil, ShaileshGargade, Vahida Attar, *Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction*,IJCSI, Vol. 9, No 3, 2012.

[8] Shivnath Ghosh, santanukoley, *Machine Learning for Soil fertility and Plant Nutrients Management using Back Propagation Neural Networks*, International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 2, Issues 2, pp.2014.

[9] NikhitaAwasthi, Abhay Bansal, *Application of Data Mining Classification Techniques on Soil Data Using R*, Vol. 4, Issues 1, pp.33-37, 2017.

[10] B.V.RamaKrishna, Dr B.Satyanarayana, *Agriculture Soil Test Report Data Mining for Cultivation Advisory*, International Journal of Computer Application (2250-1797), Vol.6, No.2, pp.11-16, 2016.

[11] Ramya M.C, Lokesh V, Manjunath T.N, Ravindra S. Hegadi, *A Predictive Model Construction for Mulberry Crop Productivity*, ICACTA, Procedia Computer Science 45, pp.156-165, 2015.

[12] B. Murugesakumar, K.anandakumar, A.bharathi, *a survey on soil classification methods using data mining techniques*, International Journal of Current Trends in Engineering & Research (IJCTER), Vol. 2 Issue 7, pp. 43 – 47, 2016.

[13] Han J and Kamber M, "*Data Mining: concepts and Techniques*", San Francisco, Morgan Kaufmann, 2001.

[14] Bhargavi, P. and Jyothi, S., *Soil classification using GATREE*. International journal of computer science and information Technology, Vol.2, No.5, pp.184-191, 2010.

[15] Dr. S.Hari Ganesh, Mrs. Jayasudha, *Data Mining Technique to Predict the Accuracy of the Soil Fertility*, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 7, pp.330 – 333, 2015.

[16] R.S. Walse , G.D. Kurundkar , P. U. Bhalchandra, *A Review: Design and Development of Novel Techniques for Clustering and Classification of Data*, International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Special Issue.1, pp.19-22, 2018.

[17] V. Parashar, *Use of ICT in Agriculture*, International Journal of Scientific Research in Network Security and Communication,Vol-4, Issue-5, pp.8-11 ,2016.

Author's Profile

Mrs.R. Jayalakshmi pursed Master of Science in Computer Science from St.Joseph's College, Trichy. She has qualified in SET eligibility examinations. Now, she is pursuing her Ph. D (Part time) in Periyar University, Salem and working as a Assistant Professor in Department of Computer Science, Sri Vidya Mandir Arts and Science College, Uthangarai, Krishnagiri, Tamilnadu, India, since 2008. Her areas of interest are Data Mining, Software Engineering, Data structures, Big Data Analytics and Operating Systems. She has 10 years of teaching experience.

Dr.M. Savitha Devi pursed Ph.D in Computer Science from Mother Teresa Women's University, Kodaikanal. She is currently working as Head cum Assistant Professor in Department of Computer Science, Periyar University Constituent College of Arts and Science, Harur, Dharmapuri, Tamilnadu, India. Her areas of interest are Digital Image Processing, Computer Networks, Data structures, and Operating Systems. She has 15 years of teaching experience and 2 years of Research Experience.

