

# Machine Learning Techniques to Predict Defects by using Testing Parameters



Prasanth Yalla, Venkata Naresh Mandhala, Valavala Abhishiktha, Chitturi Saisree, Kandepi Manogna

**Abstract:** *Since ages, the software development plays a very crucial role in the arena of software engineering. An important part here is to believe that Artificial Intelligence and Machine Learning also started its way. In the process, several metrics were analyzed, composed and some predictions were made. These predictions are very much useful to analyze the defects based on machine learning. This can be done by using various system test parameters. We found certain techniques which are used to estimate the defects based on various aspects. These features are retrieved right from the inception of the software development. In this project, we present an advance view on wide variety of Machine Learning approaches, along with different capable areas of the defects by taking their parameters.*

**Keywords:** Defect Prediction, Machine Learning, Defects, Metrics, Accuracy.

## I. INTRODUCTION

In Software Development Lifecycle (SDLC) testing plays a major role, which helps to improve the quality, efficiency, performance and reliability of the system. Software Quality is the process of scaling how the product is developed and ensuring the level of Quality to satisfy users by its performance. Some of the Software Quality attributes are Correctness, Reliability, Scalability, Efficiency, absence of bugs and Testability. In the process of testing, defects are identified, which affects software quality. But Testing is time consuming and censorious phase in software life cycle [1]. Affecting the Software Quality leads to delayed timelines, cost overruns and higher maintenance costs. The traditional

approach of software quality management can be abstracted in three ways which are: considering defects of a product to decide its major causes, take off the underlying origin of the defects and to fix them by using advanced techniques. [17].

Defect prediction is a part of testing process which is the cost-efficient action and is necessary in Software development phase. The main intent of managing defects is to fulfill customer satisfaction. Prediction is a big challenge in testing projects. So, in testing phase Defect Prediction models improve the efficiency and help developers to calculate the defect prone areas and quality in their software product [4].

According to the gatherings from many research works on predicting the defects exhibits that on an average, the Machine Learning based defect prediction models can find around 75% of defects in a product, whereas manual code can find 35-60% of defects. Machine Learning is a technique which is used to give the flexibility to the computers [23]. The definition of Machine Learning can be interpreted as, A computer program is a learn from an experience E with respect to some tasks T and performance P, the performance at tasks T, measured by P, enhance with experience E [3]. However, Machine learning is an automated data processing system with set of rules, and the system learn from the incoming data [21]. These days Machine Learning has made a tremendous change in Information Technology. Machine Learning algorithms were introduced to handle the real-world problems in a customized way. They are categorized into two types, Unsupervised and Supervised Machine Learning algorithms. Some of the Machine Learning algorithms are Naive Bayes classifier, Random Forest, Logistic Regression, Decision Tree, Multilayer perceptron etc. To evaluate the performance of Machine Learning algorithms various metrics can be used, which are known as Performance metrics. Various performance metrics to evaluate prediction for classification are confusion matrix, classification accuracy, classification report, AUC curve etc. [18]. Software Quality can be improved by identifying the defect prone modules in earlier phases of SDLC. These Machine Learning algorithms helps to categorize Non-Defect prone modules and Defect prone modules. Defect prone modules are identified and are given higher priority in testing phase of SDLC. These classifiers can also be used to identify and classify the unknown datasets by identifying the class labels. Thus, Machine Learning based approaches are applied to simplify the Defect Prediction activity. We will discuss about the various ways of study on predicting defects and algorithms used in other sections.

Manuscript published on November 30, 2019.

\* Correspondence Author

**Prasanth Yalla\***, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India. Email: [prasanthya@kluniversity.in](mailto:prasanthya@kluniversity.in)

**Venkata Naresh Mandhala**, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India. Email: [mvnaresh.mca@gmail.com](mailto:mvnaresh.mca@gmail.com)

**Valavala Abhishiktha**, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India. Email: [abhishiktha1234@gmail.com](mailto:abhishiktha1234@gmail.com)

**Chitturi Saisree**, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India. Email: [saisreechitturi1234@gmail.com](mailto:saisreechitturi1234@gmail.com)

**Kandepi Manogna**, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, A.P., India. Email: [kandepimanu@gmail.com](mailto:kandepimanu@gmail.com)

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

**II. RESEARCH METHODOLOGY**

Various studies are done on some well-known algorithms in Machine Learning. Study is done for comparative analysis research. Following are some algorithms with brief description:

**A. Naive Bayes:**

Naïve Bayes is not complex but powerful algorithm for prediction model. It is a collection of classification algorithms for multi-class and two class classification problems. This procedure is easy to learn when labeled using categorical or binary input values. It is a probabilistic classifier which classifies data by using Bayes’ theorem. Classification takes place by assuming that features are independent of each other.

**B. Logistic Regression:**

It is a statistical method where classification of dataset is done when there are one or more independent variables. The classification result is the value of one of two possible outcomes.

**C. Multilayer Perceptron:**

A perceptron is a Linear classifier which produces a single output using various real-valued inputs and their weights. Whereas, multilayer perceptron consists of more than one perceptron. In addition, multilayer perceptron is standouts among the most broadly actualized neural system topologies [22].It is made up of three layers, which are input, hidden and output layers. Input layer receives the signal, hidden layer transforms inputs into useful data for output layer. Prediction of inputs is done by output layer. This model can predict the values of unknown data by training the data with known labels.

**D. Decision Tree:**

It is a tree structured Machine Learning algorithm which is generated from training data. It comprises of leaf nodes, internal nodes and a root node. Classification starts from the root node and by testing the attributes branching takes place until the decision reaches the leaf node this classification process iterates [15].

**E. Random Forest:**

Random Forest is a feature-based selection model to predict software defects. It has the ability to perform

classification and regression techniques by using multiple decision trees to predict the accurate value. While constructing a tree significant features are randomly identified and creates an uncorrelated forest of trees.

**F. K-Nearest Neighbor:**

K-Nearest Neighbor algorithm is a statistical classification approach which can perform both regression and classification techniques. Nearest Neighbors are found in the training data and sort by the distance of all the data in training instances and these nearest neighbors and classifies unknown instance in the category of its nearest neighbor [16].

**III. LITERATURE SURVEY**

A critical survey is conducted on how Defect can be predicted using different approaches and techniques which are specified in TABLE I. The area of Testing is a crucial phase for the development of a product, but it is censorious and time-consuming phase [1]. Hence to reduce the cost and time consumption defect prediction in early stages is introduced. It improves the efficiency of testing phase and helps developers to evaluate the quality of a software product effectively [4]. Unreliability of data increases as the size of project increases which impacts on the True Positive rate of Defect prediction process [13]. To improve the software quality, Prediction of software defects metrics play a significant role in building a defect prediction model to predict software defects efficiently. Various Code metrics like Object Oriented(OO) metric, Lines of code (LOC) metric, chidambar and kemerer(CK) metrics measures the code complexity, size of code and process metrics evaluate the time and number of changes in the code during the development process which helps to predict defects efficiently [9]. Machine Learning learns automatically from training data and classifies the data into smaller form which can be used in software development phases this reduces cost and time-consumption [5]. Data is classified as Defective and Non-Defective by using software defect prediction metrics through which training instances are obtained. These instances are helpful to build a defect prediction model. Various Machine learning algorithms were used to classify and preprocess the data. From our comparative study Linear regression gives the highest prediction accuracy among some algorithms[3,4].

**Table I: Summary of survey on Research works**

S.NO	Title of the paper	Name of the Author/Year	Techniques/ Tools Used	Findings
1.	Analysis of Software Defect Prediction by Using Machine Learning Algorithms [3].	Praman Deep Singh, Anuradha Chug (2017)	Decision trees, BBN, ANN, Linear classifier, KEEL tool, WEKA tool	It is analyzed that Neural Networks have lowest error rate compared to Decision Trees. As per the result, Linear Classifier algorithm has highest defect prediction accuracy, hence it is proved to be the most reliable technique among supervised learning algorithms.

2	A Software Testing Defect Prediction Model-A novel and Practical Approach [4].	Shaik nafeez umar (2013)	Multiple Linear Regression	It is found that there is a strong association between some test parameters and defects. It is concluded that by this model there is a good and accurate prediction of defects including quality improvement. Finally, by using this model there is a probability of finding 84% of number of defectives.
3	A Progress on approaches to predict software defects[9].	Zhiqiang Li, Xiao Yuan Jing, Xiaoke Zhu (2018)	System parameters like: Codemetrics, LOC,OO metrics.	We analyzed about the model which was introduced by using, public datasets, common software metrics, defect prediction process and evaluation measures. It is concluded that this model can recognize the possible defect affecting areas in a way to achieve the quality assurance that can strongly give out less resources for code inspection and testing..
4	A Machine Learning approach to predict the software bugs[10]	Mr.Awni Hammouri, Mr.Mustafa Hammad, Mr.Mohammad Alnabhan, FatimaAlsarayrah (2018)	Decision tree, Naive Bayes, ANN, WEKA 3.6.9	It is analyzed that various metrics and performance measures were used to evaluate machine learning algorithms for bug prediction. It is concluded that by using three Machine Learning techniques (NB, DT, ANN) they resulted that these are efficient for bug prediction and Decision tree classifier gives best outcome among these.
5	A theoretical Review Study on predicting Software Defect by using Machine Learning Techniques. [5]	FeiduAkmel, ErmiyasBirihanu, Bahir Siraj (2017)	Supervised learning, Unsupervised learning, Software metrics.	A profound analysis is done on the major factors for software failures. It is concluded that machine learning algorithms makes the defect process simple, less time consuming and cost efficient compared to software metrics.
6	A Framework for Predicting Defects by using Neural Networks [20].	Mr.Vipul Vashisht, Mr.Manohar Lal, Mr.G.Sureshchandar (2015)	Neural Networks.	Through the survey multi-layer perceptron technique gave the accurate outcome compared to the other techniques like regression trees and random forest classification, BLR. In the analysis we found that Back propagation optimization is used to train the network which attained around 90% accuracy.

7	A Novel Ensemble Feature Selection and Software Defect Detection Model on Promise Defect Datasets [13].	E. Sreedevi, Y. Prasanth (2019)	SVM, Bayesian Models, Decision Tree, Neural Networks, Ensemble Models	Unreliability of data increases as the size of project increases which impacts on the True Positive rate of Defect prediction process. Dynamic multi-software ensemble classification model is developed to resolve the issue related to the software defect prediction. Various algorithms in machine learning are used to check the performance of this model and concluded that ensemble model has high defect detection rate compared to other traditional models.
8	The hierarchical Grouping of defects in software by using the decision tree algorithm [15].	M. Surendranaidu, Dr. N. Geethanjali (2013)	Decision Tree, NETBEANS (7.2), Pattern mining	It is studied that defects were categorized on attribute values: program length, difficulty, volume, effort and time estimated. Defects were classified by using ID3 algorithm. It is concluded that pattern mining technique is used after classification of defects.
9	Using Software Quality Methods to Prevent Defects and Reduce cost[17].	Rick Spiewak, Karen McRitchie (2008)	Automated tool, SEER-SEM model	The traditional method for quality management can be abstracted in three steps which are: considering the defects of products to decide root causes, alter the processes to address and take off the underlying origin of the defects and to fix them by using advanced techniques. By using Modeling tools cost effectiveness of the practices should be calculated and then they should be enforced in software construction.
10	Feature Space Transformation Technique for Predicting Software Defects [19].	Mr.Md. Habibur Rahman, Sheikh Muhammad Sarwar, Sadia Sharmin, Mohammad Shoyaib (2016)	Feature Space Transformation.	From the analysis it is know that feature space and its transformation are the factors responsible for successfully predicting defects using Machine learning. As per the result Feature space transformation gives the best defect prediction accuracy.

**IV. DATASETS AND EVALUATION METHODOLOGY**

The datasets which were used extensively by researchers are NASA datasets [12]. Therefore, these datasets are used for software defect prediction. Every dataset has various features and they are publicly available. Used datasets in this study are from NASA PROMISE Repository datasets which are KC1, MC2, MW1, PC1, PC3, PC4, PC5, PC30. These datasets measurably consist of 45 software metrics which includes 25 product metrics, 15 process metrics and 5 execution metrics [13]. Datasets were pre-processed by using various ML algorithms are explained [III].

To assess performance of Machine Learning algorithms in predicting defects, set of measures were used to calculate Accuracy from the generated Confusion Matrixes.

**A. Confusion Matrix:**

It is a performance metric, which is used in Machine

Learning and problems on Statistical classification Confusion Matrix is used. It is a table used to calculate Accuracy and effectiveness of Algorithms. It provides the report on total number of True Negative(TN), False Negative(FN), True Positive(TP) and False Positive(FP), [10].

**B. Accuracy:**

The proportion of entities to the actual known data to see if the specified algorithm predicts a True Positive. Accuracy of 1 indicates a perfect accuracy whereas 0 indicates random guess.[14]

$$Accuracy = (TP+TN)/(FP+FN+TP+TN).$$

**V. RESULT & ANALYSIS**

From our research and analysis, we selected WEKA 3.8 tool and we performed the experiment, as WEKA software contains a collection of algorithms and Visualization tools for predictive study and data pre-processing.

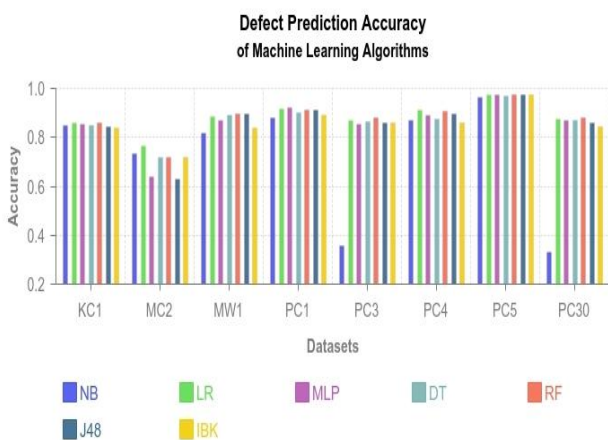




The datasets were obtained from NASA promise dataset repository in .arff format, supported by WEKA tool. Eight datasets which are KC1, MC2, MW1, PC1, PC3, PC4, PC5, PC30 were used. Then the accuracy was calculated from the obtained confusion matrix. Summary of test results were shown in defect prediction accuracy table (TABLE III). It depicts the defect prediction accuracy of each algorithm percentage-wise. The algorithm having highest accuracy in dataset is marked in bold to indicate it amongst others.

**TABLE II: Defect prediction accuracy of each algorithm.**

Algorithms	Datasets							
	KC1	MC2	MW1	PC1	PC3	PC4	PC5	PC30
Naive Bayes	84.80%	73.20%	81.80%	88.20%	35.80%	86.90%	96.30%	33.00%
Logistic Regression Classifier	85.70%	<b>76.30%</b>	88.60%	91.60%	87.00%	<b>90.90%</b>	97.27%	87.40%
Multi Layer perceptron	85.40%	63.70%	87.10%	<b>92.00%</b>	85.60%	89.20%	97.22%	87.00%
Decision table	84.70%	71.60%	89.00%	89.99%	86.60%	87.20%	97.00%	87.10%
Random Forest	<b>86%</b>	71.60%	<b>89.70%</b>	90.90%	<b>88.00%</b>	90.85%	<b>97.60%</b>	<b>88.10%</b>
J48	84.10%	62.90%	89.30%	90.90%	85.70%	89.50%	97.30%	86.00%
IBK	83.90%	71.60%	83.70%	89.19%	85.80%	86.00%	97.30%	84.10%



**Fig. 1 Defect Prediction Accuracy Chart**

From the experimental results among the algorithms Multilayer Perceptron, Naive Bayes, K-Nearest Neighbor, Decision Tree, Logistic Regression resulted that Random Forest gives the highest accuracy to predict future defects accurately, See Figure 1. However, Naive Bayes gives the lowest accuracy to predict defects.

**VI. CONCLUSION**

Various Algorithms in machine learning which are generally used are taken to predict defects. Experimental results are considered [V] and accuracy of algorithms for predicting defects which are evaluated using NASA Promise datasets [IV]. Results disclose that Machine Learning algorithms are efficient to predict defects. Through the

comparison of obtained results Random Forest classifier has the best results among other algorithms which are considered. Therefore, Machine Learning approach furnishes a better performance than other approaches which is concluded from the survey conducted and obtained experimental results. In Future, We accommodate by extending this work using Feature Space transformation technique and ensemble classification to predict defects more accurately. Furthermore, we will analyze by considering more datasets to improve efficiency.

**REFERENCES**

1. Sutar, Shantanu, Rajesh Kumar, Sriram Pai, and B. R. Shwetha. "Defect Prediction based on Machine Learning using System Test Parameters." In 2019 Amity International Conference on Artificial Intelligence (AICAI), pp. 134-139. IEEE, 2019.
2. Arora, Ishani, Vivek Tatarwal, and Anju Saha. "Open issues in software defect prediction." Procedia Computer Science 46 (2015): 906-912.
3. Singh, Praman Deep, and Anuradha Chug. "Software defect prediction analysis using machine learning algorithms." In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 775-781. IEEE, 2017.
4. Umar, Shaik Nafeez. "Software Testing Defect Prediction model-a Practical Approach." International Journal of Research in Engineering and Technology 2, no. 5 (2013): 741-745.
5. Akmel, Feidu, Ermiyas Birihanu, and Bahir Siraj. "A literature review study of software defect pre-diction using machine learning techniques." Int. J. Emerg. Res. Manag. Technol 6, no. 6 (2017): 300-306.
6. Punitha, K., and S. Chitra. "Software defect prediction using software metrics-A survey." In 2013 International Conference on Information Communication and Embedded Systems (ICICES), pp. 555-558. IEEE, 2013.
7. Jacob, Shomona Gracia. "Improved random forest algorithm for software defect prediction through data mining techniques." International Journal of Computer Applications 117, no. 23 (2015).
8. Bezerra, Miguel ER, Adriano LI Oliveiray, and Paulo JL Adeodatoz. "Predicting software defects: A cost-sensitive approach." In 2011 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2515-2522. IEEE, 2011.
9. Li, Zhiqiang, Xiao-Yuan Jing, and Xiaoke Zhu. "Progress on approaches to software defect prediction." IET Software 12, no. 3 (2018): 161-175.
10. Hammouri, Awni, Mustafa Hammad, Mohammad Alnabhan, and Fatima Alsarayah. "Software bug prediction using machine learning approach." IJACSA) International Journal of Advanced Computer Science and Applications 9, no. 2 (2018).
11. Kumaresh, Sakthi, and R. Baskaran. "Defect analysis and prevention for software process quality improvement." International Journal of Computer Applications 8, no. 7 (2010): 42-47.
12. Hall, Tracy, Sarah Beecham, David Bowes, David Gray, and Steve Counsell. "A systematic literature review on fault prediction performance in software engineering." IEEE Transactions on Software Engineering 38, no. 6 (2011): 1276-1304.
13. Mrs. E. Sreedevi, Y. Prasanth, "A Novel Ensemble Defect Detection Models For Uncertain Data", International Journal of Pure and Applied Mathematics, Vol.115, No.8, 2017, pp. 233-238.
14. <https://stats.stackexchange.com/questions/143079/what-is-prediction-a-accuracy-auc-and-how-is-it-the-number-conducted-in-machi>
15. Naidu, M. Surendra, and N. Geethanjali. "Classification of defects in software using decision tree algorithm." International Journal of Engineering Science and Technology 5, no. 6 (2013): 1332.
16. Ibrahim, Dyana Rashid, Rawan Ghnemat, and Amjad Hudaib. "Software Defect Prediction using Feature Selection and Random Forest Algorithm." In 2017 International Conference on New Trends in Computing Sciences (ICTCS), pp. 252-257. IEEE, 2017.
17. Spiewak, Rick, and Karen McRitchie. "Using software quality methods to reduce cost and prevent defects." Journal of Software Engineering and Technology (2008): 23-27.
18. T. Vijay Muni, G Sai Sri Vidya, N Rini Susan, "Dynamic Modeling of Hybrid Power System with MPPT under Fast Varying of Solar Radiation", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 1 (2017), pp.:530-537.



19. M Srikanth, T. Vijay Muni, M Vishnu Vardhan, D Somesh, "Design and Simulation of PV-Wind Hybrid Energy System", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 04-Special Issue, 2018, pp: 999-1005
20. S Ilahi, M Ramaiah, T Vijay Muni, K Naidu, " Study the Performance of Solar PV Array under Partial Shadow using DC- DC Converter", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 04-Special Issue, 2018, pp: 1006-1014.
21. S Moulali, T Vijay Muni, Y Balasubrahmanyam, S Kesav,"A Flying Capacitor Multilevel Topology for PV System with APOD and POD Pulse Width Modulation", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 02-Special Issue, 2018, pp: 96-101.
22. T Vijay Muni, S V N L Lalitha, "Fast Acting MPPT Controller for Solar PV with Energy Management for DC Microgrid", International Journal of Engineering and Advanced Technology (IJEAT), Volume 8, Issue 5, pp-1539-1544.
23. T Vijay Muni, S V N L Lalitha, "Power Management Strategy in Solar PV System with Battery Protection Scheme", International Journal of Innovative Technology and Exploring Engineering, Volume 8, Issue 6, pp-960-964.
24. [https://www.tutorialspoint.com/machine\\_learning\\_with\\_python/machine\\_learning\\_algorithms\\_performance\\_metrics.htm](https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm)
25. Rahman, Md Habibur, Sadia Sharmin, Sheikh Muhammad Sarwar, and Mohammad Shoyaib. "Software defect prediction using feature space transformation." In Proceedings of the International Conference on Internet of things and Cloud Computing, p. 72. ACM, 2016.
26. Vashisht, Vipul, Manohar Lal, G. S. Sureshchandar, and Suraj Kanya. "A framework for software defect prediction using neural networks." Journal of Software Engineering and Applications 8, no. 8 (2015): 384.
27. Krishna Mohan, G. Yoshitha, N.Lavanya, MLN.Krishna Priya, "Assessment and analysis of software reliability using machine learning techniques". International Journal of Engineering and Technology(UAE), 7(2.32 Special Issue 32), 201-205, 2018.
28. Vidya Sagar. P, Nageswara Rao .M, Venkata Naresh Mandhala, "Probabilistic estimation of software development effort techniques using machine learning". International Journal of Engineering and Technology(UAE), 7, 1085-1090, 2018.
29. Prasanth.Y, Grace Hepsiba .M, Mounika .T , Pavan Kumar. T, Raghavendra Kumar. G, "Application of machine learning techniques on naval and telecommunication system failure data". Journal of Advanced Research in Dynamical and Control Systems, Vol. 9, Sp- 6 / 2017.

### AUTHORS PROFILE



**Prasanth Yalla** received his B.Tech Degree from Acharya Nagarjuna University, Guntur (Dist), India in 2001, M.Tech degree in Computer Science and Engineering from Acharya Nagarjuna University in 2004, and received his Ph.D. degree in CSE titled "A Generic Framework to identify and execute functional test cases for services based on Web Service Description Language" from Acharya Nagarjuna University, Guntur (Dist), India in April 2013. He was an associate professor, with Department of Information Science and Technology in KL University, from 2004 to 2010. Later he worked as Associate professor, with the department of Freshman Engineering from 2011 in KL University. Presently he is working as Professor in the department of Computer Science & Engineering in KL University and also Associate Dean (R&D) looking after the faculty publications. Till now he has published 28 papers in various international journals and 4 papers in conferences. His research interests include Software Engineering, Web services and SOA. He taught several subjects like Multimedia technologies, Distributed Systems, Advanced Software Engineering, Object Oriented Analysis and design, C programming, Object-Oriented programming with C++, Operating Systems , Database management systems, UML etc. He is the Life member of CSI and received "Active Participation- Young Member" Award on 13-12-13 from CSI. He has applied a project to SERB very recently.