I J C T A, 9(20), 2016, pp. 59-67

Survey on Software Fault Detection Based on Data Mining Techniques

Divya Khurana* and Anand Prakash Shukla**

ABSTRACT

Nowadays, various faults may occur in a system during developing phase through which the state of system failure takes place that are not affordable by a developer. A software system passes through various tests then made available to the user but user also face many problems that may arises due to presence of faults in the system. There are numerous fault detection tools and techniques are made available for developer, tester as well as user but in case of complex software system it becomes a crucial as well as expensive problem. Therefore, we need to identify these faults from the system and recover that system. This paper presents the basic concept of fault detection and its approaches and also software based fault detection techniques. Moreover, paper explores the performance evaluations that how fault detection can be done by using Data mining tools (Classification algorithms) for a system or in a communication network. Furthermore, this paper explains the basic concept of Classification algorithm.

Keywords: fault detection, software based fault detection techniques, software based fault detection techniques and classification algorithms.

1. INTRODUCTION

Fault detection is the process of analysing the system in order to find faults and its location in it. Detecting faults in a system is a time consuming process, it is also based on input data. It is a subfield of control engineering and problem in process engineering. In general terms, Fault detection is a phenomenon of detecting failure in a check. There are four co-related terms [16][17] are as follows:

Faults: A fault is a condition that fails a system to perform its particular task accurately.

Failure: Failure causes when system not able to perform its specification.

Error: The difference between actual output and predicted output.

Bugs: When an expected or incorrect result is produced cause due to error, failure and faults.

Fault causes problem in system through which system is incapable of fulfilling its complete specification. The root cause of fault need to be found out in order to complete the system specifications. In the development phase, a number of faults may arise in the system which causes failure of a system. So, we need to discover these unpredictable faults in a system for the purpose of removing it. As all know that there is a fault tolerance mechanism is inbuilt in the system in the process of developing phase of system through which system can handle numerous faults that are occur in it. In the fault tolerance mechanism, it allows a system to perform its task properly in the case of failure of any one of its component. But some faults which are not being removed from the system during testing that are needed to be discovered. In developing phases, Fault can be detected through fault tolerance mechanism and also removed by it. The faults that are not removed during testing can be detected by some of the techniques are as follows:

* Krishna Institute of Engineering and Technology, Ghaziabad, India, *Email: diva.btechcs@gmail.com*

** Krishna Institute of Engineering and Technology, Ghaziabad, India, *Email: ap.shukla@kiet.edu*

- **Fault seeding**: Fault seeding [18] is the process of introducing one or more faults in the system in order to detect other existing faults in the system. It is used to enhance the effectiveness of the test-cases.
- **Mutation Testing**: Mutation testing [21] is the process used to design new test cases and then estimate the quality of existing tests. It includes some small modification in a program.
- **Fault Injection**: Fault injection [3][21] is the process of introducing one or more faults in the system in order to assess the location of existing faults and behaviour of the system in the presence of faults. It is used for improving the test coverage of the System.

Fault Detection in software is not an easy task to do. For software quality estimation, faults can be predicted using software metrics based on quality classification algorithms or fault prediction models. However, in order to build a quality estimation models is a difficult task because noisy and faulty data usually degrade the performance and quality of the system. Therefore, we need to improve the system quality and performance based on classification algorithms.

Classification [19] is the act of categorising things into some form of classes. It is one of the pervasive problems that encompass many distinctly dissimilar applications. Classification task is the collection of records. The basic concept of classification can be done by following steps:

Here, an input set of attributes (sample data set) can be classified into a particular output class label or a category. There are many classification algorithms are proposed earlier, some of which popular algorithms are: Decision Tree based Methods, Rule-based Methods, Memory based reasoning, Neural Networks, Naïve Bayes and Bayesian Belief Networks, Support Vector Machines.

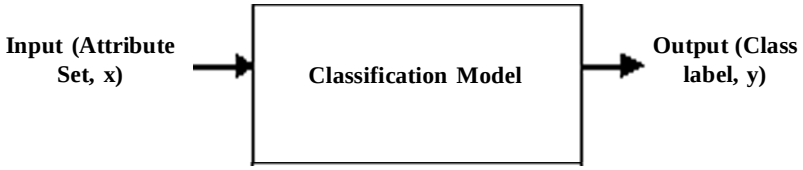


Figure 1: Classification as mapping an input attribute x into its class label y.

This paper discussed that how classification algorithms are used for detecting the faults in a system in order to improve system quality and performance. Furthermore, paper describes the mechanisms of fault detection. The main purpose of paper is to discover faults and noisy data from the system that degrades the quality and performance of the system and delay to fulfil the required specification of the system.

2. FAULT DETECTION

Fault detection [11] is the monitoring system for identifying the faults that has been occurred in the system. This method includes that all the faults that has been occurred in the system are identified and remove from the system to remove hindrance from the system to meet its required specifications. There are numerous faults that are occurred in the system it can be either software fault or hardware faults that become hindrance in the system target specification. In all the system, there is an inbuilt fault tolerance mechanism that can handle all system faults and remove from the system. But, some faults cannot detect and removed by fault tolerance mechanisms [8] which are need to be removed. The process of fault detection acquires three steps are as follows:

1. Identification of all faults that occur in the system.
2. Isolating each faults with others and its location.
3. Recovery of the system.

One of the method that are used to indicate faults i.e. sensor reading of pattern recognition. After discovery of faults, Isolated each faults with each other and categorise the type faults and its locations.

Fault detection and isolation can be categorised into two parts which are as follows:

1. Model based FDI: The Model-based FDI technique is used to decide about the occurrence of faults in the system. The system model may be knowledge based or mathematical. There is one method that comes under Model based FDI known as set-membership methods. This method works under the set of conditions and guarantee the detection of faults from the system.
2. Signal Processing based FDI: The signal processing FDI techniques is used to perform some mathematical or statistical operations on measurements in order to extract information about the faults that occurred in the system.

2.1. Fault Detection Approaches

Fault detect provide some approaches to detect and isolate faults that are occurred in the system. Following are Approaches of fault detection [10] to detect and isolate faults that are as follows:

1. **Model based reasoning:** When the output experimental system of models used as origin for fault detection and diagnosis that is called Model based Reasoning.
2. **Fault signatures, pattern recognition and classifiers:** Pattern recognition is simple approach that directly uses the output symptoms and compares it with the set of known symptoms for each possible fault. The term 'Pattern' and 'Fault signature' can be represented as a vector of symptoms for each defined faults. The loop hole of several pattern recognition approaches is that inherently make a Single fault assumption. Some of its forms are used in Event oriented diagnostics.
3. **Neural networks**: Neural networks in information technology, it is a system programs and data structures that estimates the operation of Human Brain. Neural Networks are non-linear and multi-variable models that are created by set of input and output data sets. It can be used to detect events and trends that has been occurred in the system and referred as event detector.
4. **Procedural or workflow approaches:** Many fault detection techniques are based on procedures and flow diagrams for the purpose of identifying faults. This type of approach can be handle by generating procedures, flow diagrams etc for decision making process which is based on experimental data.
5. **Event-oriented fault detection, diagnosis and correlation:** Pre-implemented approach to detect events is neural networks. An event can be used to represent the change of state of monitored object. Event oriented fault detection is the process of highly focused in the state of object over monitoring.
6. **Passive system monitoring vs. active testing:** In this approach, during the case of online monitoring system, many detection techniques may require a routine scanning of the whole system and each and every component of the system. This is required for the purpose of maintenance which is based on testing phase.
7. **Rule-based approaches and implementations:** Rule-based approach is control based mechanism which is used in the system. This approach basically used to store and manipulate knowledge in order to retrieve useful data from it for detecting faults in the system.

2.2. Advantages And Disadvantages of Fault Detection

Fault detection techniques may yields many advantages over system. Some of which are as given follows:

1. It helps to use prediction models to assess quality risk and potential defective areas.
2. Reduce test effort based on process measures.
3. Compress testing schedule and decrease costs.
4. Maximise development speed based on skimping software quality.

Fault detection techniques may have some disadvantages that are as follows:

1. Difficulty in finding faults in large and complex system.
2. Requires the construction of a least one secondary program.

2.3. Software Based Fault Detection Techniques

Nowadays, Fault detection becomes a crucial problem for software. To identify these defects and made system available to meet all its required specifications some software based fault techniques [12] are given below:

1. **Algorithm Based Fault Tolerance (ABFT):** ABFT is a technique which is used to detect, locate, isolate faults and then recover faulty system with the help of procedures and skills. This technique is only used for limited set of problems but it is an effective technique.
2. **Assertions:** Assertions are the logic statements that are not visible to the programmer and inserted at different programming levels in order to exploits the effectiveness which is depends upon the nature of application and on programmer's ability.
3. **Control Flow Checking (CFC):** CFC is the approach in which application program code can be partitioned into basic building blocks or branch free part of codes. A number or a deterministic signature can be assigned to each building block and then fault can be detected by comparing these signatures. This technique faces the problem of test case granularity.
4. **Procedure Duplication (PD):** Procedure duplication is the technique which is used to detect duplicate procedures and then compare these procedures with each other on two different processors. This method can be done manually and might introduce some errors. It requires complete checking of output result also.
5. **Software Implemented Error Detection and Correction (EDAC):** In this technique, with the help of specific software error can be identified and correct the application program code. Some of the examples of Software Implemented EDAC are Cyclic Redundancy Checks or CRC, Hamming Codes, Bose-Chaudhuri- Hocquenghem or BCH etc, all these examples are more effective in detecting error from the system. This technique may have a limitation that is very high time redundancy when software is meant to be implemented.
6. **Periodic Memory Scrubbing:** This technique of fault detection can relies on periodic reloading of code. It is very effective for protecting the code segment of Operating system and application programs.
7. **Masking Redundancy:** This is the techniques that can be able to perform all its tasks properly and in running mode in the presence of faults. Only few processors can run this type of program code.
8. **Reconfiguration:** Reconfiguration is the technique in which the failed component can be removed from the system. When failure occur in the system then it may causes some other defects on other components of respective system. After identification of faults, all types of faults can be isolated according to its types and detect the portion of faulty area.
9. **Replication:** This technique ensures the reliability of the system but in terms of hardware cost it is very expensive. In this method, it is the act of copying or reproducing some phase of data in order

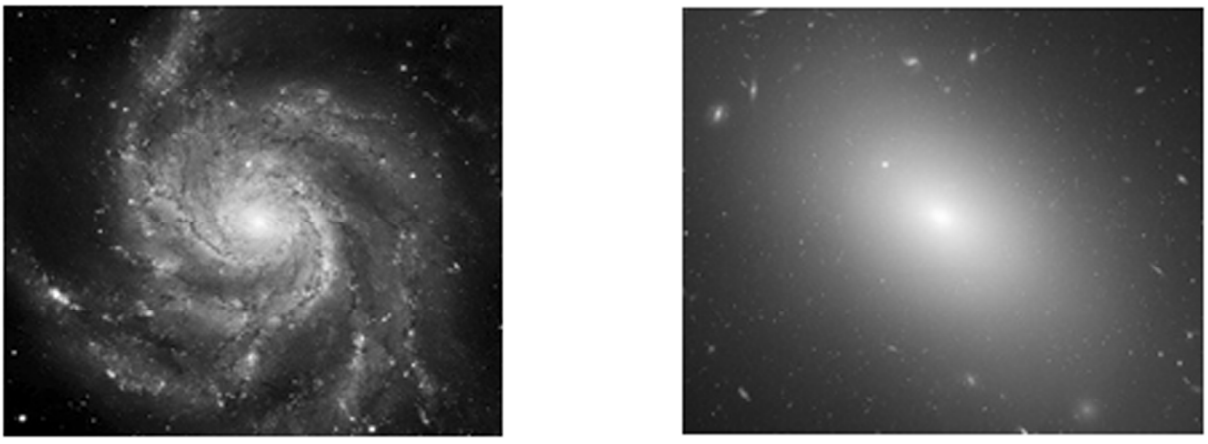
to assess fault in the system by comparing these faults with each other. This comparison is based on voting process. It slows down the computational factor.

10. **Dual Modular Redundancy (DMR) & Backward-Error Recovery (BER) & Checkpoint:** This is the technique in which error can be detected through differences in execution that is referred as Dual modular Redundancy and the process which is used to detect error in execution phase that is referred as Backward-Error Recovery. A checkpoint can be used to represent the program state and through this the snapshot of whole system programming can be taken by it.
11. **Triple Modular Redundancy (TMR) & Forward - Error Recovery (FER):** When a set of three processors working in the same environment of program then one of them suddenly fails to perform its task and lose its majority vote therefore the system can determines the erroneous throughout the processor. The Triple Modular Redundancy (TMR) is the classic example of Forward - Error Recovery (FER).
12. **Fingerprinting**: This technique of fault detection is based upon the mechanism of DMR and the execution can be done across DMR. It summaries the errors with the help of execution history in a hash-based signature.
13. **Checksum & Parity**: This technique is based on bit error rate effectiveness. It is only beneficial for error correction not for error detection. The odd number of bit error can be detected by single parity checks and remaining even number of bit errors are said to be undetected. The whole system can be implemented by typical Checksum function.
14. **Matrix Checksums:** In this technique of detecting fault from the system can be done in a matrix by using typical checksums for each row and column that are applied in a matrix. However, we can detect erroneous element of a matrix. For application program data this technique is very useful.

3. CLASSIFICATION ALGORITHMS

Classification task [19] is basically the collection of records. Each record contains the set of attributes. The main goal of classification is to predict the target program class from all records of data. It is most efficient and effective in the process of organising data into some specific categories. It makes the data more essential and made data easy to find and retrieved. In machine Learning, Classification is a technique which is used to predict group membership for data instances.

For Example: Classification of Galaxies based upon their shapes are as given below:



(a) **(b)**

Figure 2: (a) Spiral Galaxy (b) An Elliptical Galaxy

The implementation [9] of classification task can be done as follows in Figure 2. This implementation takes following steps that are given below:

1. The classification task can be applied to a set of attributes or record data set that can be named as Training set. Those sets of data are needed to be trained with the help of training algorithms.
2. The training set of data then induced after applying learning algorithms in it and start its training by learn models. After that the trained data set create a Model for showing its capabilities.
3. Trained model can be applied to a apply model for showing its ability of doing task and made available in an open environment. So, after deduction that trained set is converted into a test set of relevant data.

The classification algorithms [14] can be classified on the basics of two categories:

1. **Supervised Learning**: It includes a set of training examples. It is a machine learning task which is used to inferring a function from labelled training data. Algorithms that are comes under Supervised learning such as Artificial Neural Network, Decision Tree, Bayesian statistics, Navie Bayes Classifier, Support Vector Machine, Backpropogation etc
2. **Unsupervised Learning**: It is also a type of machine learning which is used to draw inferences from datasets that includes the input data without labelled responses. Algorithms that are come under unsupervised learning such as Clustering, Neural Networks, anomaly detection etc.

Some of the popular Classification Learning algorithms are describes briefly below:

1. **Clustering:** In this Algorithm, a trained input data set is applied to the system. Clustering algorithm is an unsupervised learning algorithm which is used to partitions a dataset into subsets (clusters) such that data in each subset are similar (according to some distance measure).

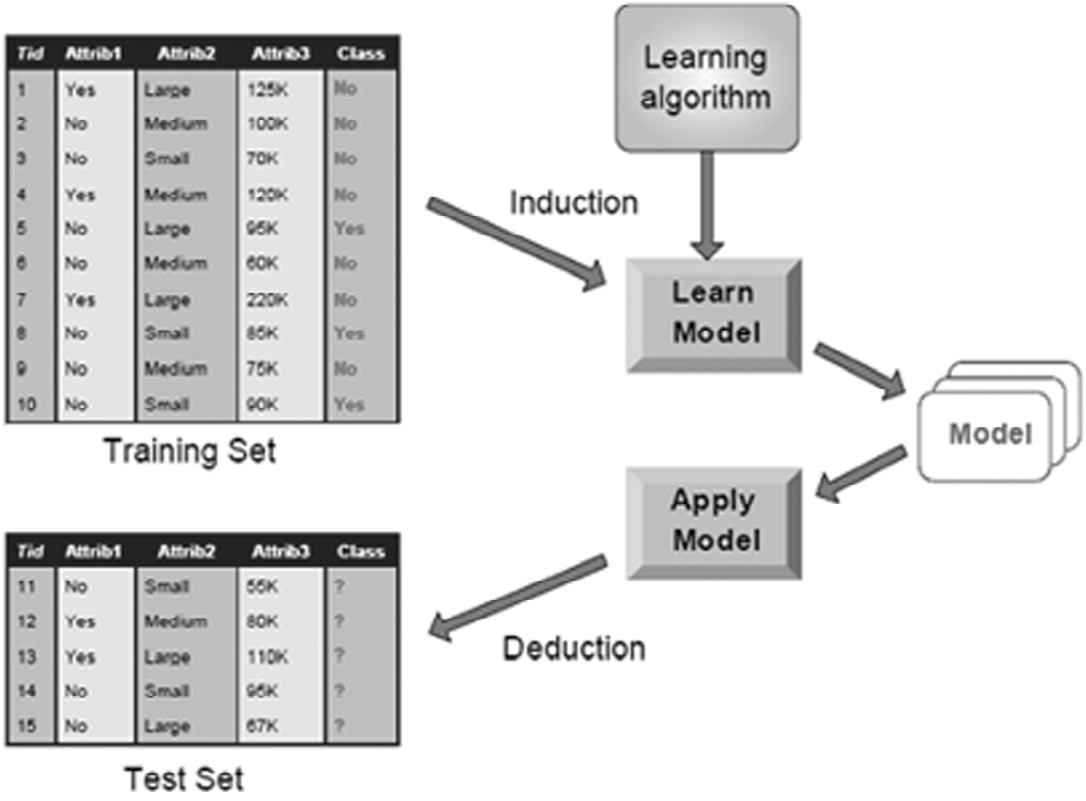


Figure 3: Implementation of Classification Task

2. **Decision Tree**: A decision tree is a supervised machine learning model which is predictive in nature. It decides the target value of new samples which is based on available attributes data set values. A decision tree is a hierarchical set of questions that are used to classify an element.
3. **Support Vector Machine (SVM):** Support Vector Machines are supervised learning algorithm which is used for both classification as well as regression. SVM algorithm can be applied to both linear and non-Linear data. The data on which this algorithm needs t be binary. SVM make use of certain Kernel in order to transform the problem. Kernel equations may be linear, quadratic, Gaussian etc.
4. **Neural Network:** Neural Network is an unsupervised learning algorithm in which a computer system can be represented on the human brain and human nervous system. It includes a large number of processors set which are operating in parallel and each processor has its own small sphere of knowledge.

4. PERFORMANCE EVALUATION FOR FAULT DETECTION USING DATA MINING TECHNIQUES

Author Name	*Project name*	*Problem Statement*	*Conclusion*
Golriz Amooee, Behrouz Minaei-Bidgoli, Malihe Bagheri-Dehnavi	A Comparison Between Data Mining Prediction Algorithmsfor Fault Detection [2]	Many maintenance activities have a goal to reduce failure of industrial machinery and also decrease the occurrence of these failures. Therefore, many companies try to improve their efficiency by using distinct fault detection techniques.	The main aim of this author is to make use of data mining tools for discovery of defective component of the system. So, Firstly, remarkable points are needed to find out which are related to the defective components. After that, with the help of integrated database, identifying outliers, clean up the data and ignore constant variables that are occurred and then apply different prediction algorithms. This paper also includes the purpose to improve industrial product's reliability, maintainability and thus availability.
Alfonso Capozzoli, Fiorella Lauro, Imran Khan	Fault detection analysis using data mining techniques for a cluster of smart office buildings [4]	Nowadays, a rapid growth of automated fault detection tools is required. For the purpose of reducing abnormal consumption of the system. A number of methods are required to handle this problem but it is a very complex problem and cannot be deal by using other equipments. In this paper the basic approach for detecting automatic anomalies in forming energy consumption which is based on actual recorded data of active electrical power for lighting and total active electrical power of a cluster.	In this research can be aimed at testing that are based on Data mining techniques and ANN BEM in combination for the purpose of automated fault detection process. By reducing numerous fault anomalies for creating new application in order to improve the fault detection process. Anomalies. This Paper also proposed and implemented for different potential and limitations to have proven inadequate for the purpose of detection.
Poonam Chaudhary & Vikram Singh	A Data Mining Tool for Network Fault Detection [1]	System may suffer a general network fault management problem, namely, fault detection, isolation, and diagnosis has been taken up in this communication.	This paper was proposed and implemented in the WEKA environment. This paper proposed the result of the comparison of two Classification Algorithm namely, J48 and MLP and with also one proposed model. All techniques have dame data set from Ericson. The

(*contd...*)

(*Table 1 contd...*)

Author Name	*Project name*	*Problem Statement*	*Conclusion*
			basic concepts of k-mean clustering, neural network training and J4.5 classification have been used in the proposed approach model.
Jyoti Tamak	A Review of Fault Detection Techniques to Detect Faults and Improve the Reliability in Web Applications [22]	Nowadays, a major concern is reliability of software in Industry. Most of the software shows some bugs in the software after released. So, detection of these faults in a large-scale will become difficult. In order to estimate the software reliability of the software, earlier we use Software reliability growth models (SRGM). This paper presents the Research in the field of software reliability techniques for the purpose of locating bugs in the system. System may show more complex behaviour in web-based applications then. Therefore, the concern regarding software reliability in web based applications will need to be explored over different kinds of networks.	This paper focuses on growth of software reliability models. In this paper, the useful information in order to improve the system reliability. This survey show how SRGM used to calculate the time delay and minimise the cost of software system and the large-scale application for critical issues.

5. CONCLUSION AND FUTURE WORK

Faulty system gives an unreliable and insufficient environment to user. There are numerous faults occurred in the system during software development phase that cannot be handled by inbuilt software fault tolerance mechanism. Therefore, fault detection becomes an important task. There are numerous fault detection tools and techniques are already proposed in order to handle these faults and for system recovery process but they are very expensive and require highly efficient environment.

In this paper, we have described a basic concept of fault detection and in order to detect faults by inexpensive manner. We have also mentioned the software based fault detection techniques. Moreover, For the purpose of fast and accurate output of system we have described the performance evaluation of software fault detection by using Data Mining tools (Classification algorithms such as J48 Decision tree, SVM etc.)

This paper also described the term “Classification” and shows how classification algorithm become useful for fault detection Mechanism. Here, we observed in many research that both supervised and unsupervised algorithms gives a best results for identifying faults.

Future works in this field may helpful to achieve quicker and accurate results and prevent failure. Also focus on frequency of failure during fault detection mechanism.

REFERENCES

[1] Poonam Chaudhary, Vikram Singh,” A Data Mining Tool for Network Fault Detection”, Ijcsc, Vol 6 , Number 2, April - Sep 2015 pp, 275-280.

[2] Golriz Amooee, Behrouz Minaei-Bidgoli, Malihe Bagheri-Dehnavi’, “A Comparison between Data Mining Prediction Algorithms”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[3] Roberto Natella, Domenico Cotroneo, Joao A. Duraes, and Henrique S. Madeira, “On Fault Representativeness of Software Fault Injection”, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. 1, JANUARY 2013.

[4] Alfonso Capozzoli, Fiorella Lauro, Imran Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings", Article in press, No. of Pages 15, Model 5G, ESWA 9790, 24 January 2015.

[5] Joao A. Duraes, and Henrique S. Madeira, "Emulation of Software Faults: A Field Data Study and a Practical Approach", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 32, NO. 11, NOVEMBER 2006.

[6] Poonam Chaudhary, Vikram Singh, "NETWORK FAULT DETECTION - A CASE FOR DATA MINING", International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166, Volume 5 Issue , 4 July 2014.

[7] A. Jefferson, J. Huffman Hayes, "A Semantic Model of Program Faults", ISSE-TR-95-110, August 1995.

[8] Brian Randell, "System Structure for Software Fault Tolerance", IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL SE-1, NO. 2, JUNE 1975.

[9] Tan, Steinbach, Kumar, " Data Mining Classification: Basic Concepts, Decision Trees, and Model Evaluation", Lecture Notes for Chapter 4, April 2004.

[10] http://gregstanleyandassociates.com/whitepapers/FaultDiagnosis/faultdiagnosis.htm.

[11] Janos Gertler, "Fault Detection and Diagnosis in Engineering Systems", Basic concepts with simple examples.

[12] Jyoti Tamak, "A Review of Fault Detection Techniques to Detect Faults and Improve the Reliability in Web Applications", Volume 3, Issue 6, June 2013.

[13] Rob Schapire, "Machine Learning Algorithms for Classification", Princeton University.

[14] http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm.

[15] Michael R. Lyu, ' Software Reliability Engineering: A Roadmap', Future of Software Engineering (FOSE '07). IEEE Computer Society, Washington, DC, USA, 153-170, 2007.

[16] Haissam Ziade ; Rafic Ayoubi2 ; Raoul Velazco, A Survey on Fault Injection Techniques, The International Arab Journal of Information Technology, Vol. 1, No. 2, July 2004.

[17] www.softwaretestingtimes.com/2010/04/fault-error-failure.html?m=1.

[18] A Pasquini, E De Agostino, "Fault seeding for software reliability model validation", Volume 3, Issue 7, July 1995, Pages 993-999.

[19] http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

[20] http://www.tutorialspoint.com/software_testing_dictionary/mutation_testing.htm

[21] Regina Lúcia de Oliveira Moraes and Eliane Martins, 'Fault injection approach based on architectural dependencies' , In Architecting Dependable Systems III, Rogério Lemos, Cristina Gacek, and Alexander Romanovsky (Eds.). Springer-Verlag, Berlin, Heidelberg 300-321.

[22] Jyoti Tamak, "A Review of Fault Detection Techniques to Detect Faults and Improve the Reliability in Web Applications", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013