

# Statistical Anomaly Detection Technique for Real Time Datasets

**Y.A.Siva Prasad**

Research Scholar, Dept of CSE  
KLUniversity, Andhra Pradesh

**Dr.G.Rama Krishna**

Professor, Dept of CSE  
KLUniversity, Andhra Pradesh

## Abstract:

Data mining is the technique of discovering patterns among data to analyze patterns or decision making predictions. Anomaly detection is the technique of identifying occurrences that deviate immensely from the large amount of data samples. Advances in computing generates large amount of data from different sources, which is very difficult to apply machine learning techniques due to existence of anomalies in the data. Among data mining techniques, anomaly detection plays an important role. The identified rules or patterns from the data mining techniques can be utilized for scientific discovery, business decision making, or future prediction. Several algorithms has been proposed to solve problems in anomaly detection, usually these problems are solved using a distance metric, data mining techniques, statistical techniques etc. But existing algorithms doesn't give optimal solution to detect anomaly objects in the heterogeneous datasets. This paper presents statistical control chart approach to solve anomaly detection problem in continuous datasets. Experimental results shows that proposed approach give better results on continuous datasets but doesn't perform well in heterogeneous datasets.

**Keywords – Outlier, Data Mining, Patterns.**

## I. INTRODUCTION

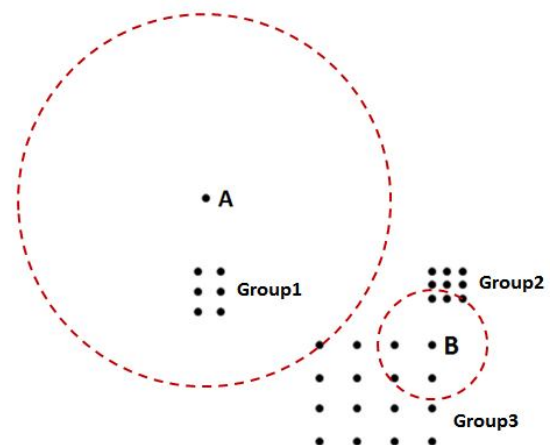
Anomaly detection is categorized into two subcategories: supervised, or unsupervised. Supervised Anomaly detection determines the class of an observation from the classifiers. The classifiers are classified as the machine learning models whose parameters are learned from the training dataset. The primary challenge in constructing the classifiers is the skewness in the dataset between the normal class and anomaly. Due to the reason that the training data for anomalies is too small relatively based on the data for the normal class. . In the absence of training data for the Anomalies, the semi supervised methods may be used. In this instance, a machine learning model is chosen to capture the boundaries of the normal class. A new observation that falls outside of the boundaries is classified as an anomaly.

Anomaly detection techniques effort to find the objects that might be different from the rest of the data objects in a given data set. Usually, anomalies are generated from

certain misbehavior of the data points which get very different from the majority of data set. Random samples which are greatly deviates from its neighbors in relation to its local compactness is treated as outlier. The compactness is measured via the length of the k-nearest neighbor distances of its neighbors. Although a local outlier might not exactly differ from all other observations.

Statistical approaches are the standard algorithms applied to outlier detection. The main aim of these approaches is that normal data objects follow a generating mechanism and abnormal objects deviate through generating mechanism. Given a certain type of statistical distribution , algorithms compute the parameters assuming all data points are generated by such a distribution (mean and standard deviation). Outliers are points that possess a low probability to be generated by the entire distribution (deviate greater than 3 times the standard deviation from the mean). These methods have the limitation that they will assume the data distribution . Another limitation of the statistical methods is that they don't scale well to large datasets or datasets of large attributes.

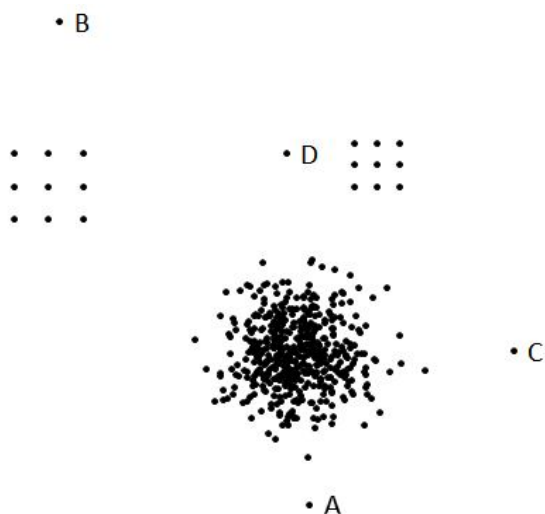
## Density Based Anomaly Detection:



Red dash circles contain the k nearest neighborhood of 'A' and 'B' when k value is 7. Point outliers is defined as Separate data objects that are different with respect to the rest of dataset.

Collective outliers is defined as "A set of data objects is considered as anomalous with respect to the entire data set, then members of the set are called collective outliers".

**Point based Anomaly detection:**



Data objects "A, B, C and D" in the figure are considered to be as point anomalies as they aren't in any cluster and are also distant from the majority of the data objects. So that they contrast from the majority of data objects by using the locality aspect. Point detection has attracted much attention in practical applications for instance fraud detection, detecting criminal activities and network attack detection.

**II. LITERATURE SURVEY**

In order to overcome the deficiencies of distance-based methods, Breunig et al. [1] proposed that each data point of the given data set should really be assigned a degree of outlier. With their view, for example other recent studies, a data point's measure of anomaly should be measured relative to its neighbors; hence they refer to it just like the "local outlier factor" of the data point. Tang et al. [2-4] argued that any outlier doesn't always have to remain of lower density and lower density isn't a necessary condition to remain an outlier. They modified LOF to search for the "connectivity-based outlier factor" (COF) which they argued is so much more effective each time a cluster and a neighboring outlier have similar neighborhood densities. Local density is widely measured in terms of k nearest neighbors; LOF and COF both exploit properties associated with k nearest neighbors of causing given object in the data set. However, it is possible that any outlier lies in a location between objects given by a sparse as well as a denser cluster. To account for such possibilities, Jin et al. [5-6] proposed another modification, called INFLO, that's in accordance to a symmetric neighborhood relationship. That is, their proposed modification considers neighbors and 'reverse neighbors' associated with a data point when estimating its density distribution. Tao and Pi [2] have proposed a density-based clustering and outlier detection (DBCOD) algorithm, which also is a member of the density-based algorithms. Density-based algorithms

assume that all neighborhoods associated with a data point have similar density. If some neighbors of one's point can be found a single cluster, plus the other neighbors near each other another cluster and to discover the two clusters have different densities, then comparing the density of a given data point with all of that neighbors may lead to a wrong conclusion and the recognition of real outliers may fail.

Anomalous series detection and contextual abnormal subsequence detection are both viewed as applicable for time series data set. In the current research, only real-valued time series are actually, categorical-valued time series are out of the coverage of this investigation. Anomalous series detection only places focus on identifying anomalous series whereas the contextual abnormal subsequence problem requires that we all detect abnormal subsequence in the context of a single series, and this requires the comparison between subsequences and the majority of this game's series. The main gap between these two techniques is: the first one works to find out which series is anomalous as the latter one wants to know when abnormal behaviors occur.

Some of the problems in Time series data are: Historical information associated with a series has to be examined, but how to summarize the useful historical details are a difficult problem. The behavior of outliers is different for different applications, and it makes detecting abnormal behavior a hard activity. Within a single application domain, the outlier is likewise changing with time, so it requires any effective algorithms or techniques to be very adaptive and malleable to contend with dynamic detection. The algorithm ought to be aware of the dynamically changing outliers[7].

Knorr et al. [3] proposed the DB(pt, dt) Outlier detection scheme, wherein an object obj is said to be to get an outlier if at the very least fraction pt of the total objects have greater than dt distance to obj. They defined several techniques to find such objects. For instance the index based approach computes distance range using spatial index structure and excludes an object if its dt-neighbourhood contains greater than 1-pt fraction of total objects. They proposed nested loop algorithm to avoid the cost of building an index. They additionally proposed growing a grid so that any two objects beginning with the same grid cell have a distance of the most dt to one another. In this way objects ought to be in relation to those from neighboring cells to examine if they're outliers.

In statistics, regression analysis is made use of to approximate the relationship between attributes. Linear regression and logistic regression are two common models. An outlier in regression analysis is undoubtedly an observation whose value is removed from the prediction. To detect such outliers, the residuals of the observations are computed dependent on a trained model.

Traditional clustering techniques effort to segment data by grouping related attributes in uniquely defined

clusters. Each data point within the sample space is granted to just 1 cluster. K-means algorithm and also its different variations will be the most well-known and commonly used partitioning methods. The value ‘k’ stands for the number of cluster seeds initially provided for the algorithm. This algorithm takes the input parameter ‘k’ and partitions a set of m objects into k clusters [4]. The procedure work by computing the gap between an information point and to discover the cluster center to enhance one item into one of this very clusters ensuring that intra-cluster similarity is high but inter-cluster similarity is low. A common method to obtain the distance will be to calculate to sum of the squared difference as shown below and it is known as the Euclidian distance.

$$d_k = \sum_n \left\| X_j^k - C_{i_j} \right\|^2$$

where,  $d_k$  : is the distance of the  $k^{\text{th}}$  data point from  $C_j$

### III. PROPOSED SYSTEM

When mining the different real-time data using pattern mining algorithms, it needs to preprocess the data, and then receives training dataset by analyzing the features of data used in anomaly detection. If we have good algorithms, and not high-quality training data, the detective result will be not good. Different from other application fields, anomaly detection usually uses some artificial intelligence methods, which analyze data by choosing a model. However choosing models always depends on instinct and expert knowledge, and there isn't an objective method to evaluate the data.

In this proposed approach a statistical control chart algorithm is used in order to find the anomalies in different continuous datasets. We proposed dynamic 3 sigma based control chart to detect anomalies in the dataset. Basic flow structure of the proposed algorithm is shown below:

#### Algorithm:

Input: Continuous dataset

Output: Dataset without anomalies.

Procedure:

Step 1: Load dataset with continuous attributes.

Step 2: Check each attribute in the dataset as real attribute or not.

Step 3: Calculate mean and standard deviation of each attribute.

Step 4: Calculate upper control limit of each attribute (2)

Step 5: Calculate control limit of each attribute (3).

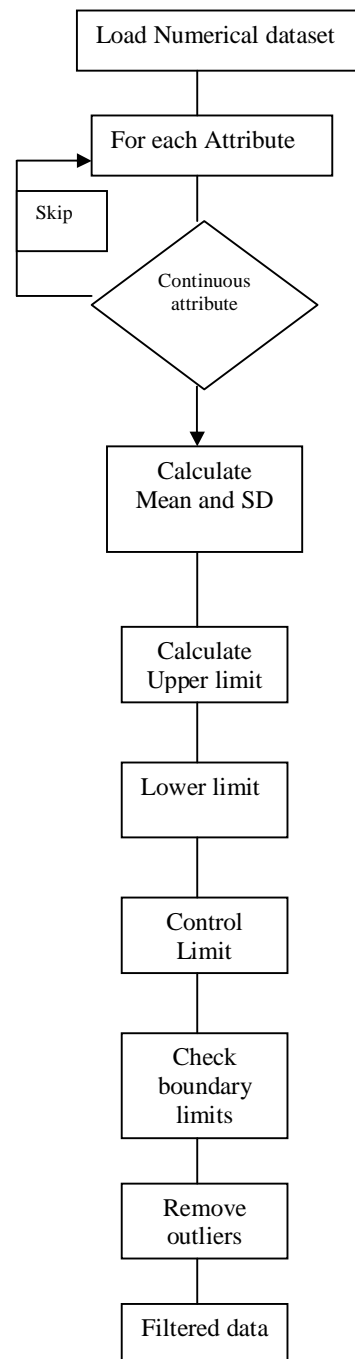
Step 6: Calculate lower control limit of each attribute (1)

Step 7: Check whether each object in the dataset falls within three categories i.e lower, upper or control limits.

Step 8: If the object is out of bound then it is removed from the dataset instances.

Step 9: This process is repeated until all data points are completed.

Step 10: Finally dataset without outliers are stored in file.



Flow chart of Proposed algorithm

Lower Control limit:  $\mu_x - \lambda\sigma_x$  --(1)

Upper Control Limit:  $\mu_x + \lambda\sigma_x$  ---(2)

Control Limit:  $\mu_x$  ----(3)

In order to check the performance of this algorithm a decision tree algorithm c4.5 algorithm is used to check the accuracy of the original dataset with filtered dataset.

**C4.5 Summarized algorithm:**

C4.5 is an improved version of decision trees over ID3 from the training data, using the concept of information entropy. The training data is a set  $s_1, s_2, s_3 \dots$  represents data objects in the dataset  $S$ . Each  $s_i = x_1, x_2, \dots$  is a sample values where  $x_1, x_2, \dots$  represents features or attributes of the sample. The training data associated with a vector  $C = c_1, c_2, \dots, c_n$  where  $c_1, c_2, \dots, c_n$  represents the class to which each sample belongs to dataset. Every node of the decision tree, C4.5 chooses one attribute of the data the most efficiently splits its range of samples into subsets in a single class. Its criterion will be the improved information gain that outcomes from selecting an attribute for splitting the data. The attribute with the highest calculated information gain is selected to get the decision attribute. The C4.5 algorithm then recurs on the smaller sublists. This algorithm has got a few base cases. All of the samples within the list remain in the very same class. When that happens, it basically causes a leaf node for the decision tree same to select that class. · Not one of the features provide any information gain. In this instance, C4.5 causes a decision node higher up the tree utilizing the expected value of the class.

Modified Information or entropy is given as

$$\text{ModInfo}(D) = -S_i \sum_{i=1}^m l \log \sqrt{S_i} \text{ , } m \text{ different classes}$$

$$\begin{aligned} \text{ModInfo}(D) &= -S_1 \sum_{i=1}^2 l \log \sqrt{S_i} \\ &= -S_1 \log \sqrt{S_1} + S_2 \log \sqrt{S_2} \end{aligned}$$

Where  $S_1$  indicates set of samples which belongs to target class ‘anomaly’,  $S_2$  indicates set of samples which belongs to target class ‘normal’.

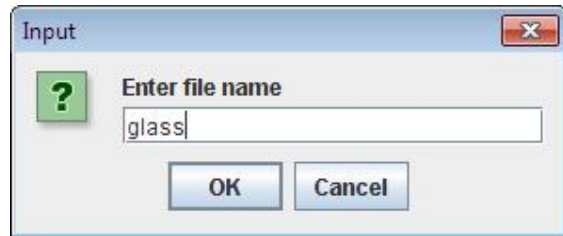
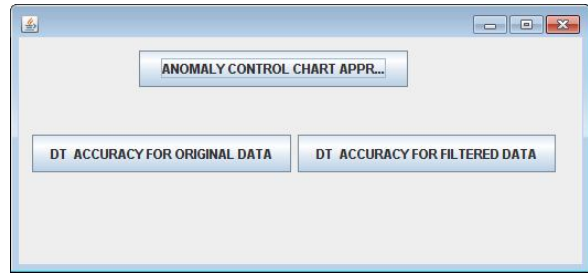
Information or Entropy to each attribute is calculated using

$$\text{Info}_A(D) = \sum_{i=1}^v |D_i| / |D| \times \text{ModInfo}(D_i)$$

The term  $D_i / |D|$  acts as the weight of the  $j$ th partition.  $\text{ModInfo}(D)$  is the expected information required to classify a tuple from  $D$  based on the partitioning by  $A$ .

**IV EXPERIMENTAL RESULTS**

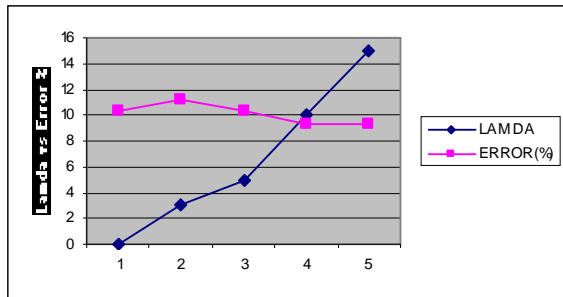
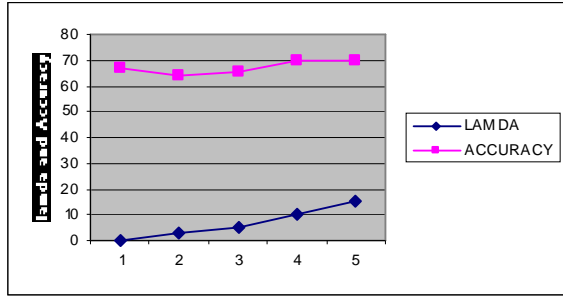
All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operating system platform is Microsoft Windows XP Professional (SP2).



Min outlier and max outlier 1.2146790466222046  
 1.8220517944992913  
 Min outlier and max outlier -68.25250510420858  
 95.06820603878802  
 Min outlier and max outlier -141.55625177676382  
 146.92531719732455  
 Min outlier and max outlier -48.48205801799239  
 51.371871102104535  
 Min outlier and max outlier -4.803644897071962  
 150.10551405595052  
 Min outlier and max outlier -64.72212848113165  
 65.71624063066434  
 Min outlier and max outlier -133.35838611131706  
 151.27231134496193  
 Min outlier and max outlier -49.54687933099827  
 49.8969727889422  
 Min outlier and max outlier -9.686860717855696  
 9.800879409444482  
 Non Outliers :163  
 Total Outliers :51  
 Total DataSet Size :214

**Glass dataset:**

| Lamda           | 0       | 3       | 5       | 10      | 15      |
|-----------------|---------|---------|---------|---------|---------|
| <b>Error</b>    | 10.26   | 11.15   | 10.35   | 9.33    | 9.33    |
| <b>Outliers</b> | 0       | 36      | 49      | 51      | 51      |
| <b>Accuracy</b> | 66.8224 | 64.0449 | 65.4545 | 69.9387 | 69.9387 |



**Diabetes Dataset:**

Min outlier and max outlier -13.002838230160977  
 20.692942396827643  
 Min outlier and max outlier -38.968559725681104  
 280.7576222256811  
 Min outlier and max outlier -27.67356710322389  
 165.8845046032239  
 Min outlier and max outlier -59.22462950530506  
 100.29754617197172  
 Min outlier and max outlier -496.4205325900252  
 656.0194909233585  
 Min outlier and max outlier -7.428223476877225  
 71.41337972687718  
 Min outlier and max outlier -1.1847666729805415  
 2.128519277147207  
 Min outlier and max outlier -25.560272286726736  
 92.04204312006007  
 Non Outliers :724  
 Min outlier and max outlier -13.002838230160977  
 20.692942396827643  
 Min outlier and max outlier -38.968559725681104  
 280.7576222256811  
 Min outlier and max outlier -27.67356710322389  
 165.8845046032239  
 Min outlier and max outlier -59.22462950530506  
 100.29754617197172  
 Min outlier and max outlier -496.4205325900252  
 656.0194909233585  
 Min outlier and max outlier -7.428223476877225  
 71.41337972687718  
 Min outlier and max outlier -1.1847666729805415  
 2.128519277147207  
 Min outlier and max outlier -25.560272286726736  
 92.04204312006007  
 Non Outliers :725

Total Outliers :43

**C4.5 decision rules:**

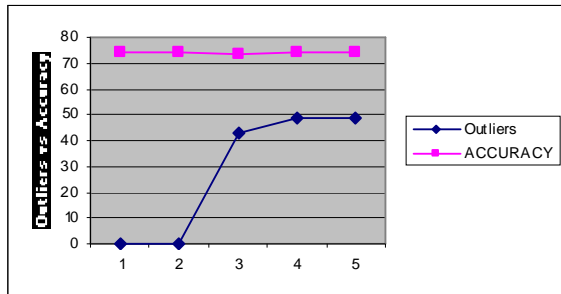
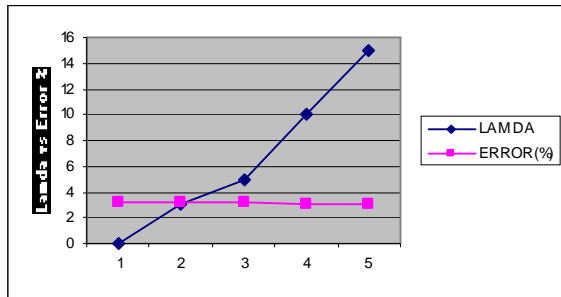
```

plas <= 143
| mass <= 26.3: tested_negative (136.0/4.0)
| mass > 26.3
| | age <= 28: tested_negative (211.0/33.0)
| | age > 28
| | | plas <= 100
| | | | insu <= 152
| | | | | preg <= 3: tested_negative (14.0)
| | | | | preg > 3
| | | | | plas <= 0: tested_positive (2.0)
| | | | | plas > 0
| | | | | | pedi <= 0.787: tested_negative (33.0/3.0)
| | | | | | pedi > 0.787: tested_positive (4.0/1.0)
| | | | | insu > 152
| | | | | | skin <= 27: tested_positive (3.0)
| | | | | | skin > 27: tested_negative (3.0/1.0)
| | | plas > 100
| | | | age <= 58
| | | | | pedi <= 0.527: tested_negative (95.0/43.0)
| | | | | pedi > 0.527: tested_positive (46.0/11.0)
| | | | age > 58: tested_negative (12.0/1.0)
plas > 143
| plas <= 154
| | age <= 42
| | | preg <= 7: tested_negative (30.0/9.0)
| | | preg > 7: tested_positive (5.0/1.0)
| | age > 42
| | | pedi <= 0.251
| | | | insu <= 50: tested_positive (3.0/1.0)
| | | | insu > 50: tested_negative (2.0)
| | | | | pedi > 0.251: tested_positive (11.0)
| plas > 154
| | mass <= 29.8
| | | age <= 61
| | | | age <= 25: tested_negative (3.0)
| | | | age > 25
| | | | | mass <= 27: tested_positive (10.0)
| | | | | mass > 27: tested_negative (4.0/1.0)
| | | | age > 61: tested_negative (4.0)
| | mass > 29.8: tested_positive (88.0/11.0)
    
```

Number of Leaves : 21

Size of the tree : 41

| Lamda           | 0       | 3       | 5       | 10      | 15      |
|-----------------|---------|---------|---------|---------|---------|
| <b>Error</b>    | 3.158   | 3.158   | 3.208   | 3.094   | 3.094   |
| <b>Outliers</b> | 0       | 0       | 43      | 49      | 49      |
| <b>Accuracy</b> | 73.8281 | 73.8281 | 73.1034 | 74.4089 | 74.4089 |



## V. CONCLUSION AND FUTURE SCOPE

This paper presents the Statistical anomaly detection process for real time continuous datasets. The experiment was conducted based on glass, diabetes datasets. The experiment has shown which datasets have anomaly, has produced the anomaly detection patterns which can be used to remove the outliers in the datasets. Proposed approach doesn't give optimal solution to heterogeneous datasets. Proposed approach fails to identify the anomalies when the lamda value is low. Experimental shows proposed work on different datasets by varying lamda values. Results shows proposed approach works well on the continuous attributes.

## REFERENCES

- [1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying densitybased local outliers," In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM Press, pp. 93–104, 2000.
- [2] Y. Tao and D. Pi, "Unifying density-based clustering and outlier detection," 2009 Second International Workshop on Knowledge Discovery and Data Mining, Paris, France, pp. 644–647, 2009.
- [3] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *VLDB*, 1998.
- [4] Pavel Berkhin, "Survey of Clustering Data Mining Techniques", <http://citeseer.ist.psu.edu/berkhin02survey.html>
- [5] K. P. Chan and A.W. C. Fu, "Efficient time series matching by wavelets," In Proceeding ICDE '99 Proceedings of the 15th International Conference on Data

Engineering, Sydney, Australia, March 23-26, 1999, p. 126, 1999..

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, p. ARTICLE 15, July 2009.

[7] R. Fujimaki, T. Yairi, and K. Machida, "An anomaly detection method for spacecraft using relevance vector," in *Learning, The Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, 2005, pp. 785–790.

## AUTHORS PROFILE



Mr.Y.A.Siva Prasad, Research Scholar in CSE Department,, KL University, Andhra Pradesh, and Life member of **CSI**, **IAENG**, Having 9 years of Teaching Experience and presently working as an Associate Professor at Chendhuran College of Engineering & Technology ,Pudukkotai, Tamilnadu .